ARTIFICIAL NEURAL NETWORKS BUILT FOR THE RECOGNITION OF ILLICIT AMPHETAMINES USING A CONCATENATED DATABASE*

STELUTA GOSAV, MIRELA PRAISLER

Department of Physics, "Dunărea de Jos" University, Str. Domneasca nr. 47, Galați, Romania,
E-mail: sgosav@ugal.ro

Received December 15, 2008

In this paper we are presenting several expert systems built for the identification of illicit amphetamines using GC-FTIR spectra, GC-MS spectra and a hybrid GC-FTIR - GC-MS spectral database (concatenated spectral database). The systems were built using Artificial Neural Networks (ANN), and are dedicated to the recognition of amphetamines. The database is formed by chemical compounds with toxicological relevance, representing drugs of abuse (mainly central stimulants, hallucinogens, sympathomimetic amines, narcotics and other potent analgesics), precursors and derivatized counterparts.

Key words: drugs of abuse, GC-FTIR, GC-MS, concatenated database, ANN.

1. INTRODUCTION

ANN is a branch of the field known as "Artificial Intelligence" and consists of a system loosely modelled based on the human brain. The field is referred to in many ways, such as connectionism, parallel distributed processing, neuro-computing, natural intelligent systems, machine learning algorithms, and artificial neural networks. ANNs have ability to account for any functional dependency. The network discovers (learns, models) the nature of the dependency without needing to be prompted. There is no need to postulate a model, or to amend it. ANNs have the ability to learn from experience in order to improve their performance and to adapt themselves to changes in the environment. In addition, they are able to deal with incomplete information or noisy data, and can be very effective especially in situations where it is not possible to define the rules or steps that lead to the solution of a problem [1–7].

This paper presents a comparative analysis among seven systems: the IR-ANN system which has as input variables GC-FTIR spectra, the MS-ANN

* Paper presented at the 4th National Conference on Applied Physics, September 25–26, 2008, Galați, Romania.

system which has as input variables GC-MS spectra, the IR-MS-ANN system which uses as input variables both types of spectra (GC-MS and GC-FTIR) concatenated in a single hybrid "spectrum" for each compound in the database, the 100imp_ IR-MS-ANN system which uses as input variables the 100 most important concatenated spectral data and the 100sensit_ IR-MS-ANN system which uses as input variables the 100 most sensitive concatenated spectral data. All ANN systems were designed to identify illicit amphetamines (Fig. 1) [8].

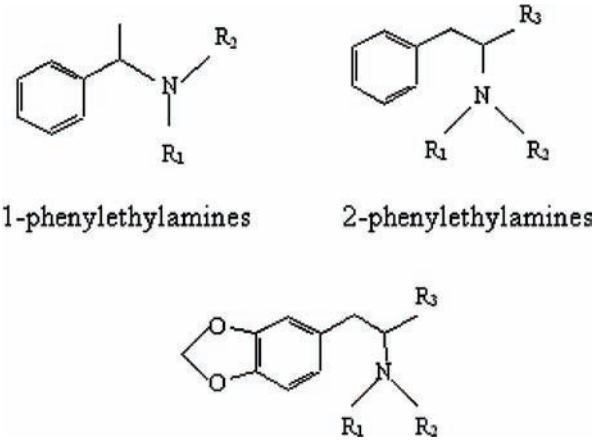

Fig. 1 – Molecular structures of the main amphetamines analogues.

2. EXPERIMENTAL PART

The experimental conditions in which the GC-FTIR spectra were recorded have been presented in a previous paper [9]. The obtained reference spectra were stored in a digital library after normalization. The scan range was from 4000 to 580 cm^{-1}. All spectra in the library were reduced in size by eliminating the spectral windows where the compounds in the database have no IR absorptions. Hence, data ranged from 3745 to 2555 cm^{-1} and from 1995 to 605 cm^{-1}. The resulting 259 wavenumber intervals of 10 cm^{-1} yielded a data matrix with 159×260 entries, as the database was formed using the infrared spectra of 159 forensic substances such as drugs of abuse (mainly central stimulants, hallucinogens, sympathomimetic amines, narcotics and other potent analgesics), precursors, or derivatized counterparts. The samples represent reference standards and laboratory synthesized compounds. In our case, classification experiments were carried out on 159 spectra of 15 stimulant amphetamine analogues (class code M), 8 hallucinogenic amphetamine analogues (class code T) and 136 nonamphetamines (class code N).

The mass spectra (electron impact ionization) of the compounds included in the GC-MS database were imported from general MS libraries (NIST mass

spectral database, AAFS spectral library, and an in-house-made MS library). The spectra from all data basis were recorded in standard conditions. The MS spectra range from m/z 12 to 260. The database contains 103 mass spectra of forensic substances, out of which 15 are stimulant amphetamines, 5 hallucinogenic amphetamines and 83 nonamphetamines.

In this paper, we have built seven neural networks aiming the automatic class identity assignment for amphetamines. The architecture of all ANN systems presented in this paper consists of three layers of neurons or nodes, which are the basic computing units: the input layer, one hidden layer and the output layer. The latter has three nodes, one for each class of the modeled compounds (stimulant amphetamines – class M, hallucinogenic amphetamines – class T and nonamphetamines – class N). The back-propagation algorithm was used for training the networks. Before this process, the following network parameters have been optimized: the number of nodes in the hidden layer, the learning rates, and the momentum. We have adopted the sigmoid function as transfer function (activation) for all neural networks. The networks were built using the *Easy NN plus* software, which runs under Microsoft Windows XP and may be used on IBM-compatible personal computers.

The training set of the artificial neural networks (IR-ANN and 100IR-ANN networks) which use as input variables *only* the GC-FTIR spectra, consists of 29 samples: 7 stimulant amphetamines, 5 hallucinogenic amphetamines and 17 nonamphetamines. The remaining 130 samples were included in the validation set. The stimulants and hallucinogens included in the training set are the following: N-ethylamphetamine, amphetamine, methamphetamine, N-n-propylamphetamine, β-phenylethylamine, α-phenylethylamine and N-methyl-α-phenylethylamine, and 3,4-methylenedioxyamphetamine, 3,4-methylenedioxy-methamphetamine, 3,4-methylenedioxy-N-ethylamphetamine, N-methyl-1-(3,4-methylenedioxyphenyl)-2-butanamine and 1-(3,4-methylenedioxyphenyl)-2-butanamine. The stimulant and hallucinogenic amphetamines of the training set have been selected in order of the similarity of their spectra with those of the parent compounds (amphetamine for the stimulants and 3,4-methylenedioxy-ampetamine for the hallucinogens). The nonamphetamines included in the training set were selected randomly from the IR database: bemegride, β-butyro-lactone, cadaverine and its HFB-derivate, codeine and its pentafluoropropionic (PFP)-derivate, caffeine, γ-butyrolactone, the trimethylsilyl (TMS)-derivate of γ-hydroxy butyric acid, the TMS-derivate of γ-hydroxy valeric acid, γ-valerolac-tone, nicotamide, piracetam, putrescine, dextromoramide, nicotine and prolintane.

The neural networks (MS-ANN and 100MS-ANN networks) which use as input variables *only* the GC-MS spectra have the same training set (7 stimulants, 4 hallucinogens and 17 nonamphetamines) and validation set (75 samples). The stimulant and hallucinogenic amphetamines from the training set are: α-phenyletilamine, amphetamine, mephentermine, methamphetamine, N-ethylam-

phetamine, N-methyl-α-phenylethylamine and N-n-propylamphetamine, and 3,4-methylenedioxyamphetamine, 3,4-methylenedioxymethamphetamine, 3,4-methylenedioxy-N-ethylamphetamine and 1-(3,4-methylenedioxyphenyl)-2-butanamine. The nonamphetamines included in the training set are following: bemegride, diethylpropion, cadaverine, fencamfamine, codeine, 1-phenyl-2-propane, caffeine, γ-butyrolactone, methylphenidate, norephedrine, γ-valerolactone, nicotamide, nicotine, dextromoramide, yohimbine, morphine and lidocaine.

The training set used by the neural networks (IR-MS-ANN, 100imp_IR-MS-ANN and 100sensit_IR-MS-ANN) with input variables a concatenated database (GC-FTIR and GC-MS spectra) is formed by 7 stimulants, 4 hallucinogens and 17 nonamphetamines. The validation set contains 75 samples. The stimulant and hallucinogenic amphetamines from the training set are: α-phenyletilamine, amphetamine, methamphetamine, N-ethylamphetamine, N-methyl-α-phenylethylamine, β-phenylethylamine and N-n-propylamphetamine, and 3,4-methylenedioxyamphetamine, 3,4-methylenedioxymethamphetamine, 3,4-methylenedioxy-N-ethylamphetamine and 1-(3,4-methylenedioxyphenyl)-2-butanamine. The nonamphetamines included in the training set are following: bemegride, diethylpropion, cadaverine, fencamfamine, codeine, 1-phenyl-2-propane, caffeine, γ-butyrolactone, methylphenidate, norephedrine, γ-valerolactone, nicotamide, nicotine, dextromoramide, yohimbine, morphine and lidocaine.

The IR-ANN and MS-ANN networks have as input all the absorptions or abundancies measured in the spectra of the compounds in the database: 260 input variables representing absorption intensities measured 10 cm^{-1} apart respectively, 247 input variables representing the abundances of the fragments. The second step was to build two networks, 100IR-ANN and 100MS-ANN, which have as input only 100 variables, *i.e.* the 100 most important IR absorptions, and the 100 most important abundances respectively. The third step was to build the IR-MS-ANN network which uses a concatenated input database, *i.e.* the 100 most important IR absorptions and 247 abundances. We have choose these input variables for IR-MS-ANN system because they have the best modeling/discrimination power, *i.e.* the 100IR-ANN and MS-ANN networks which use as input variables the 100 most important IR absorptions, and the 247 abundances respectively, have the best values for the validation parameters.

In order to optimize the sample/variable ratio, to increase the data-processing speed and to improve the efficiency of the class identity assignment, we have applied two selection criteria of input variables: the importance and sensitivity criteria [10–12]. For the process optimization we have choose the best network, namely IR-MS-ANN network, which has 347 input variables (the 100 most important IR absorptions for the IR-ANN network and all abundancies – 247). The absolute importance of an input variable (input node) is done of the sum of absolute weights of the connections among this input node and the nodes of the hidden layer. The importance analysis is a method for measure the influence of

each input upon the next (hidden) layer in the network. The absolute sensitivity is a measure of how much the outputs change when the inputs are changed. The change in the outputs is measured as each input is increased from lowest to highest to establish the sensitivity to change. The sensitivity analysis is a method for measuring the cause – effect relationship between the input layer and the output layer. The fourth step was to optimize the sample/variable ratio in the case of IR-MS-ANN network using the importance and sensitivity criteria. Applying these criteria, we have obtained two new ANN systems: 100imp_IR-MS-ANN and 100sensit_IR-MS-ANN networks.

3. RESULTS AND DISCUSSIONS

In order to compare the efficiency of the networks, we have performed the validation process by using all the samples from the database. The validation method was full cross-validation, as the number of samples in the database was relatively small. In order to evaluate the performance of the ANN systems, several figures of merit for the classification were calculated: the rate of true positives (TP), of true negatives (TN), of false positives (FP), of false negatives (FN), of classification (C), and of correctly classified samples (CC). Their values are presented in Table 1.

Table 1

Validation results

	IR-ANN	100IR-ANN	MS-ANN	100MS-ANN	IR-MS-ANN	**100imp_IR-MS-ANN**	100sensit_IR-MS-ANN
TP(%)	90.9	*100*	*100*	100	100	**100**	100
TN(%)	85.5	*80.74*	*78.95*	56.95	90.09	**93.83**	88.89
FP(%)	14.5	*19.26*	*21.05*	43.07	9.1	**6.17**	11.11
FN(%)	9.1	*0.0*	*0.0*	0.0	0.0	**0.0**	0.0
C(%)	96.23	*98.74*	*97.08*	87.38	94.17	**98.05**	98.05
CC(%)	86.27	*83.44*	*84.00*	68.89	92.78	**95.05**	91.09

By comparing the figures of merit of the first four ANN systems, we can see that the selection of the input variables, according to their importance, leads to an improvement of the validation results only in case of GC-FTIR spectra. The best performing ANN system built with IR absorptions is 100IR-ANN network, and the best performing ANN system built with MS spectra is the MS-ANN network. Secondly, the table shows that both ANN systems have a very good sensitivity (TP = 100%). However, the 100IR-ANN network is more selective than the MS-ANN network (TN = 80.74%, and TN = 78.95% respectively). The CC rates have nearly the same values for the 100IR-ANN and

MS-ANN networks (CC = 83.44%, and CC = 84.00% respectively). Moreover, the 100IR-ANN network has the best value for the C rate as well.

The IR-MS-ANN network has the best performances in comparison of the previous four networks, with the exception of the C rate. The selectivity of this network is also the best, the TN rate being greater with approximately 10% than that for the 100IR-ANN network. This result is very important, as the modeling of the class of nonamphetamines is difficult, the compounds belonging to the N class having very different FTIR and MS spectra.

A first remark regarding the performances of the IR-MS-ANN, 100imp_IR-MS-ANN and 100sensit_IR-MS-ANN networks, which use as input a concatenated database, is the fact that the all expert systems have a very good sensitivity, all positive samples being correctly identifyed. In fact, all validation parameters have very good values. Accordingly, we can emphasize that the use of the concatenated input database (GC-MS and GC-FTIR spectra) is more appropriate than the homogenous input database (containing only GC-FTIR or only GC-MS data).

Analyzing the all results of validation we have observed that the 100imp_IR-MS-ANN is the best performing network. Therefore, the important concatenated input variables produce the best modeling/discrimination power of the all class identities (stimulant amphetamines – class M, hallucinogenic amphetamines – class T, and nonamphetamines – class N). The 100imp_IR-MS-ANN has a very good sensitivity (TP = 100%) and selectivity (TN = 93.83%). In addition, this network has the largest values for the C rate of classification (C = 98.05%) and for the CC rate of correct classification (CC = 95.05%).

3. CONCLUSIONS

The class identity of amphetamines can be assigned with better accuracy by using hybrid GC-FTIR - GC-MS spectral database rather than homogeneous spectral database (GC-FTIR or GC-MS spectra). The challenge which appears during the optimization process of a network is the fact that the improvement of selectivity determines, in many cases, a decrease of the sensitivity. Thus, the findings of this paper lead to very important results, as the idea of using a concatenated input database (FTIR and MS spectra) for an ANN system has lead to a significant increase of selectivity (TN = 90.9%) of the network, while keeping in the same time a very good sensitivity (TP = 100%). The ANN systems built with homogenous spectral data bases, described in our previous papers [8], have a lesser selectivity in the identification of illicit amphetamines, *i.e.* the best value of the TN rate being of 85%. In conclusion, the IR-MS-ANN network which uses the concatenated input variables generates the best efficiency in the automated recognition of the class identity of amphetamines.

In order to improve the selectivity and sensitivity of IR-MS-ANN network, the selection of the input variables must be performed. The selection criteria cut out those input variables that contribute less to the modeling/discrimination power of the classes, in order to eliminate the redundant information as well as for testing a new sample once the screening system is built. The selection of input variables (concatenated database) using the importance criterion leads to better validation results than if the sensitivity criterion is used. In conclusion, the best analytical performances has been obtained with the 100imp_IR-MS-ANN system.

The higher efficiency of the 100imp_IR-MS-ANN network seems to be generated by the fact that the modeling of the positives is correlated with the stability of the absorptions, regardless of their intensity. This seems to be true even for the spectral windows are characterized by a constant *lack* of absorptions for a given class of compounds. From an analytical point of view, GC-MS spectra bring of the 100imp_IR-MS-ANN network spectral information complementary to that offered by the GC-FTIR spectra and their combination allows a more efficient positive of the *individual* identity of an unknown sample (as opposite to the determination of the *class* identity).

REFERENCES

1. J. Zupan, J. Gasteiger, *Neural Networks for Chemists. An introduction*. VCH, Weinheim, 1993.
2. F. Despagne, D. L. Massart. *Analyst,* **123**, 157R–178R, 1998.
3. D. L Massart, B. G. Vandeginste, L. M. C. Buydens, S. De Jong, P. J. Lewi, J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics*: Part B, Elsevier, 1997.
4. M. Praisler, I. Dirinck, J. Van Bocxlaer, A. P. De Leenheer, D. L. Massart, *Anal. Chim. Acta* **404**, 303–317, 2000.
5. S. Gosav, R. Dinica, M. Praisler, *J of Mol. Struct,,* **887**, 1–3, 269–278, 2008.
6. P. N. Penchev, G. N. Andreev, K. Varmuza, *Anal. Chim. Acta* **388**, 145–159, 1999.
7. M. Praisler, S. Gosav, *Anal. Univ. 'Dunarea de Jos' Galati*, Fascicula II 97–110, 2002.
8. S. B. Karch, *Drug Abuse Handbook*, CRC Press, New York, 1998.
9. S. Gosav, M. Praisler, D. O. Dorohoi, G. Popa, *J of Mol. Struct,* **744–747**, 821–825, 2005.
10. M. H. Zhang, Q. S. Xu, F. Daeyaert, P. J. Lewi, D. L. Massart, *Anal. Chim Acta,* **544** 167–176, 2005.
11. A. Hunter, L. Kennedy, J. Henry, I. Ferguson, Comp. *Meth. and Progr in Biomed*, **62**, 11–19, 2000.
12. M. Gestal, M. P. Gomez-Carracedo, J. M. Andrade, J. Dorado, E. Fernandez, D. Prada, A. Pazos, *Anal. Chim. Acta*, **524**, 225–234, 2004.