HUMANN-Based Computational Neural System for the Determination of Pollutants using Fluorescence Measurements

C.P. Suárez Araujo[a]; J.J. Santana Rodríguez[b]; P. García Báez[c]; J.R. Betancort Rodríguez[b]

[a] Institute of Cybernetic Sciences and Technology; University of Las Palmas de G.C.; 35017 Las Palmas de G. C.; Spain; e-mail: cpsuarez@dis.ulpgc.es

[b] Dept. of Chemistry, Faculty of Marine Sciences; University of Las Palmas de G. C.; 35017 Las Palmas de G. C., Spain; e-mail: jsantana@dqui.ulpgc.es

[c] Dept. of Statistics, Operating Research and Computation; University of La Laguna; 38071 La Laguna, Spain; e-mail: pgarcia@ull.es

Keywords: Unsupervised Artificial Neural Network, HUMANN, Polychlorinated Byphenyls, Polychlorinated Dibenzofurans, Fluorescence Spectrometry

Abstract

Polychlorinated Biphenyls (PCBs) and Polychlorinated Dibenzofurans (PCDFs) are chlorinated aromatic compounds and are emitted into the environment. It has been shown to have toxicity and carcinogenic potential characteristics. Because of this their identification and quantification is a matter of great concern. However, the similar structure of PCBs and PCDFs can produce overlapping in fluorescence spectra, which add difficult to their determination.

We present in this paper, an HUMANN-based computational neural system [1][2] for the identification of these compounds. HUMANN is a multilayer neural net with high biological plausibility. Its adaptive character is essentially embodiment in the labelling module, because of its dynamic dimension The determination of the different analytes will be indicated by the firing neurons in the labelling layer and by the activation level of these neurons.

In this work its have also been developed a model for spectral data, fluorescence spectrum of single compounds and complex mixture, via Gaussian distribution.

Our final proposal consists in putting to work together fluorescence spectrometry and neural computation approach, and to analyse the good results and the troubles found in this new method using three type of spectra: excitation, emission and synchronous.

Introduction

PCBs is a family of compounds produced commercially by the direct chlorination of biphenyl. These compounds have found application in a wide variety of industrial uses due to their chemical and thermal stability. Since their discovery in environmental samples in 1966, it has been generally accepted that they are ubiquitous in every component of the global ecosystem [3].

PCDFs are chlorinated tricyclic aromatic compounds and are emitted into the environment as unwanted by-products of anthropogenic processes. PCDFs have been globally distributed and are found in all environmental media. They are chemically stable, have low solubilities in water, and have been shown to accumulate in the foodchain. It is well known that the PCDFs with chlorines substituted in the 2,3,7,8 positions are thought to pose a risk to human health due to their toxicity, carcinogenic potential and potential effects on animal reproductive and inmunological systems [4].

Taking account the importance of the environmental effects of these compounds, its determination constitute an important aspect of the control of quality of the environment, principally the marine environment.

From an analytical chemistry point of view, different techniques have been used to determine these pollutants: gas chromatography, liquid chromatography and mass spectrometry. Alternately, luminescence spectrometry has been also used for the study of PCBs [5] and PCDFs [6].

The application of luminescent techniques, specifically, fluorescence spectrometry, to the analysis of mixtures of compounds is particularly attractive due to the high sensitivity that can be achieved. However, this method has an extremely restricted scope of application in the analysis of complex

mixtures because its selectivity is reduced by the extensive spectral overlap, above in the case of compounds of chemical similar structures, like PCBs or PCDFs.

In order to face this problem we propose an HUMANN-based computational neural system for the identification of organic compounds. It has a structure of pre-processing and processing stage.

The proposed system has important advantage refering to other computational solutions based in artificial neural networks [7] because of HUMANN can perform blind clustering and it has a strong adaptive character. In addition, this neural computational method uses only spectral fluorescence data, is very simple, fast and economic method for monitoring of the environment.

Methods and Experiments

HUMANN-based computational neural system for fluorescence identification of pollutants compounds

The proposed system consists of two parts: The pre-processing module, which is the responsible to obtain a vector of characteristics for fluorescence spectra to analyse. This vector will be introduced in the processing module. This module is made up for HUMANN and it performs the determination of different analytes present in the mixture.

Pre-processing module

During the learning process the artificial neural networks create internal representations of the characteristics of the training set patterns. The objective of the pre-processing stage is to prepare the information environment of HUMANN in such a way that it can adequately extract the information required. The pre-processing stage constructs a set of feature vectors of the real time signals to be analysed from these same signals.

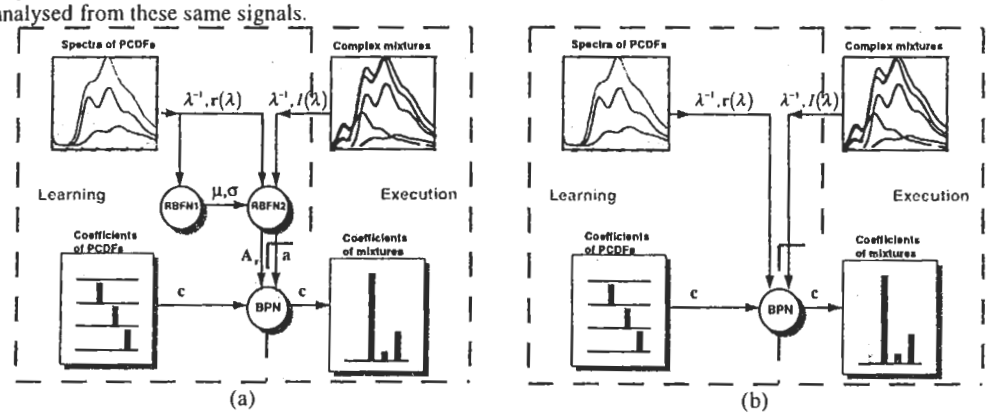


Figure 1: Scheme of pre-processing stage a) with RBFNs and b) without RBFNs

In the analysis of real fluorescence spectra two stages are followed:

1. Modelling of the spectra.
2. Determination of the features vector which corresponds to the fluorescence spectrum model of complex mixture.

Using as a base the developments and later experimental studies made by Lloyd and Evett [8] and Cabaniss [9], the fluorescence spectra can be modelled by a Gaussian distribution of intensity versus reciprocal wavelength (frequency). The emission spectrum can be represented as equation:

$$I(\lambda) \approx \sum_i a_i \exp\left(-\frac{\left(\lambda^{-1}-\mu_i^{-1}\right)^2}{2\sigma_i^{-2}}\right) \qquad (1)$$

where μ_i are the Gaussian means (cm), σ_i are the standard deviations (cm) and a_i are the amplitudes of each spectrum. If now:

$$gaus_i(\lambda^{-1}) = \exp\left(-\frac{(\lambda^{-1}-\mu^{-1})^2}{2\sigma_i^{-2}}\right) \quad (2)$$

then:

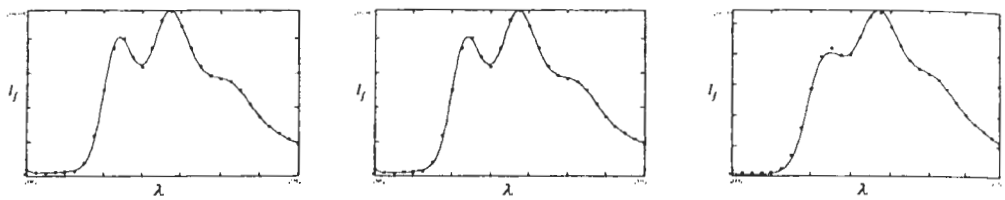
$$I(\lambda) \approx \mathbf{a} \cdot \mathbf{gaus}(\lambda^{-1}) \quad (3)$$

We will work with a linear approximation for the complex mixture model, such that the spectrum of a mixture will be represented by a linear combination of reference spectra [10]. The reference spectra are the spectra of the compounds which can be identified in a mixture.

$$I(\lambda) \approx \mathbf{c} \cdot \mathbf{r}(\lambda) = \sum_i c_i r_i(\lambda) \quad (4)$$

$$I(\lambda) = \mathbf{c} \cdot \mathbf{A}_r \cdot \mathbf{gaus}(\lambda^{-1}) \quad (5)$$

where $\mathbf{r}(\lambda)$ are the reference spectra and $\mathbf{c}$ is a vector with the contributions of each of the spectra of $\mathbf{r}(\lambda)$ in the mixture. $\mathbf{c}$ is then a vector which characterises a mixture and which is ideal for use as a vector of characteristics for HUMANN.



(a) (b) (c)

Figure 2: Fitting between model using 8 gaussians and experimental emission spectrum: a) 3DBF using equation (3), b) 4DBF, using equation (3), c) mixture 3DBF+4DBF using equation (5)

Spectral representation via Gaussian distribution will be carried out using radial basis function networks (RBFN) [11], and the approximation of concentration coefficients vector (**c**) through a backpropagation network [12]. Our pre-processing module is therefore made up of a complex neural structure of RBFNs + BPN, Figure 1.

Two RBFNs are used to obtain spectral representation. One will determine the parameters which define the Gaussian distribution and the other will approximate the intensities of the different spectra modelled. An initial radial basis function network (RBFN1) is made up of an input layer of one neuron, a hidden layer whose number of neurons coincides with the number of Gaussians which carry out the approximation of the real time spectrum (in our case 10 Gaussians) and an output layer the size of which determines the number of compounds which can be identified in any complex mixture to be analysed. This network will allow us to determine the parameters that define the Gaussian distributions which model the fluorescence spectra of each compound belonging to PCDFs and PCBs families and mixtures that can be found in real environmental samples.

The training set of the RBFN is the group of standardised spectra in reciprocal wavelength λ^{-1} of the substances, while the supervising signal is made up of the different values of spectral intensity of each simple substance. In this way, this network will converge to values of μ (vector of the Gaussian means) and σ (vector of the standard deviations) which will be used to obtain the subsequent approximations of the amplitudes **a** of the Gaussians of the various spectra of the PCDFs.

Once μ and σ have been obtained another RBFN2 is constructed in which the only layer of variable weights is the output layer, corresponding to the amplitude values **a**, with the remaining parameters set at the values μ and σ previously obtained. Copies of this network are used for training with all the available spectra, both of substances and complex mixtures. In this way, the approximations of amplitude a are obtained for the spectrum as a Gaussian function defined by the parameters μ and σ.

Once the Gaussian approximations of the fluorescence spectra have been obtained, the features vector (c) is determined. Working from the linear model of complex mixtures given by equation (4) and from the Gaussian approximation given by equation (3) , we have:

$$\mathbf{c} \cdot \mathbf{r}(\lambda) \approx \mathbf{a} \cdot \mathbf{gaus}(\lambda^{-1}) \tag{6}$$

$$\mathbf{c} \cdot \mathbf{A_r} \cdot \mathbf{gaus}(\lambda^{-1}) \approx \mathbf{a} \cdot \mathbf{gaus}(\lambda^{-1})$$

$$\mathbf{c} \cdot \mathbf{A_r} \approx \mathbf{a}$$

$$\mathbf{c} \approx \mathbf{a} \cdot \mathbf{A_r^{-1}}$$

where $\mathbf{A_r}$ is a matrix which stores the amplitude vectors resulting from the Gaussian approximation of each of the reference spectra.

In order to determine the coefficient vector **c** a three layers backpropagation network was designed with linear activation functions. The size of the input layer is given by the number of Gaussians used in the approximation of the fluorescence spectra, the output layer is made up of as many neurons as compounds to be identified, while the maximum size of the hidden layer will be the same as that of the output layer 292. The input patterns of this network will be vector **a** and the network output will be the features vector **c** which is being looked for. The training set of this network is made up of the spectra of the compounds, for which we know the concentration coefficients, and which will act as contributions vector c. In the Figure 2, it is shown the goodness of fitting between modelled and experimental fluorescence spectra using two PCDFs.

The proposed pre-processing module is an optimised module following the Gaussian representation model of fluorescence spectra. A cruder, but possible, approximation would be one which only considers the use of the linear model of mixtures in the estimation of the features vector. In the results section the validity of this Gaussian approximation will be demonstrated. Indeed it is in this particular case where the best results are obtained.

Processing Module (HUMANN)

The new hierarchical unsupervised modular adaptive neural network (HUMANN), is the processing module. HUMANN has high biological plausibility. The main causes of this are, its adaptive character, it performs self-organising processes, its modular functionality and it has connection structure with two types of synaptic connections present in the biological neural network, namely the active synapses and the silent synapses [13].

The adaptive character of HUMANN is essentially embodiment in the labelling module, because of its dynamic dimension [1][2]. This characteristic is implemented by two neuronal mechanism, a) neuronal elimination, b) neuronal generation. They perform refinement processes in the neuronal circuits, and they are present in the human brain and in the brain of some birds [14].

HUMANN will be responsible for the identification and determination of the different PCDFs and PCBs existing in complex mixture samples of environmental interest, without prior knowledge of which (and how many) of these possible compounds can be found in the analysed sample. With this purpose in mind HUMANN will be trained using the characteristic vectors **c** obtained in the pre-processing module.

HUMANN version used in this application has specific characteristics of implementation and operational structure, which allows it to adjust its performance to the real information environment. It extends its operational field beyond the clustering processes. HUMANN implements Kohonen's SOM using the scalar product in the computation of distances, and the corresponding adaptations for the Tolerance module.

During the learning process the labelling layer carries out multiple assignation functions. In other words, the input pattern will be associated as belonging to as many classes as neurons have been fired in that layer. Again, in the tolerance layer the parameter λ will have the highest possible value within of its range [1], while none of the spectra of compounds will be classified as belonging to more than one class. HUMANN will converge to a number of neurons in the output layer (labelling layer) equal to the number of compounds in the complex mixture analysed. A compound is represented by a single neuron from the labelling layer. The firing of this single neuron will indicate the presence or absence of the corresponding compound in the analysed complex mixture.

Experiment design

For the creation of the data corpus, aspects of the data that allow the efficiency of the system designed for the detection and identification of compounds in mixtures has been considered. Based on previous studies [2], we have taken into account requirements that facilitate its study using our system based on Artificial Neuronal Networks:

- Working with a sufficiently high number of compounds, which allows us to come to reliable conclusions about the results obtained.
- Establishing a group of physical parameters for the generating of spectra which avoid undesirable alterations. These parameters include excitation and emission wavelengths (λ), $\Delta\lambda$ and the wavelength interval belonging to each spectrum.
- Obtaining emission, excitation and synchronous spectra from the compounds used and from any mixtures of them. In this way we will contrast the performance of the resolution of the mixture depending on type of spectrum used in the process of resolution.
- Varying the concentration range of each compound to determine and model the alterations that took place, in the resolution, as a result.
- Saving the identifying data of the spectra tested, especially those that could explain changes in their physionomy.

These requirements will be established in the selection of some groups of spectra with particular characteristics. Some of these characteristics will define the problem to be resolved. Others are designed to avoid the destabilising effect that the variability of these characteristics could have on the shape of the spectra, which would make it difficult to obtain good results in the identifications.

The indications of the chemical experts and the analysis carried out on several studies on the peculiarities of the families to be used in the study were followed in order to determine the general characteristics of the data to be collected. These are set forth in Table 1.

	PCDFs	PCBs
Analytes	DBF, 1-DBF, 2-DBF, 3-DBF, 4-DBF, 5-DBF	Bi, Mono, Tri, Di, Penta, Hexa
Concentrations	$5x10^{-7}$M, $1x10^{-6}$M, $5x10^{-6}$M	$5x10^{-7}$M, $1x10^{-6}$M, $5x10^{-6}$M, $10x10^{-6}$M
Spectra	Emission, Synchronous	Emission, Excitation, Synchronous
Number of mixtures of an analyte with different concentrations	18 (6 analytes x 3 concentrations)	24 (6 analytes x 4 concentrations)
Spectra / mezcla pura	2 per type = 4	6 per type= 18
Spectra of mixtures of an analyte with different concentrations	36 per type = 72	144 per type = 432
Number of mixtures	150 (50 x 3 concentrations)	49
Spectra / mixture	1 per type = 2	3 per type = 9
Spectra of mixtures	132 de cada tipo = 264	147 per type = 441

Table 1: General characteristics of the data corresponding to PCDFs y PCBs

The peculiarities corresponding to the different types of spectra to be used are expressed in Table 2 for the PCDF and PCB families.

	PCDFs Spectra		PCBs Spectra		
	Excitation	**Synchronous**	**Emission**	**Excitation**	**Synchronous**
λex, $\Delta\lambda$	290nm	67nm	325nm	252nm	74nm
λ interval	[300nm-370nm]	[200nm-350nm]	[270nm-400nm]	[235nm-300nm]	[235nm-325nm]
Points per spectrum	141	301	261	131	181
λ Interval for each point	0.5nm	0.5nm	0.5nm	0.5nm	0.5nm

Table 2: Characteristics of PCDFs and PCBs spectra

Under the previous conditions, the choice of the mixtures to be carried out was made in accordance with two different philosophies depending on the family. For the PCDFs a suggestion was made to generate all of the possible mixtures that contained combinations of up to 4 analytes, where all the analytes that constituted a mixture had an equal concentration. In this way, of the combinations of 2, 3,

and 4 analytes out of a possible 6, using three different concentrations for each compound in each mixture, resulted in a total of 150 mixtures.

Among the difficulties found in this procedure, stand out the fact that a high number of mixtures (150) is needed, even taking into account that the maximum number of analytes in a mixtures was not high (4). To increase the number of analytes per mixture, it is necessary not to generate all of the possible mixtures, trying to rule out some of them so that they would influence the reliability of the final results as little as possible. To this end, a random design of the of the synthesis tables with the PCB family was established. This design was generated automatically considering that the distributions of the analytes - as they were grouped by number of analytes that each mixture contained - were as close as possible to being the same. Thus, mixtures of up to 6 analytes were achieved with the idea of generating only 49 mixtures.

From the designs of previous experiments, it was necessary to disqualify some of the samples used once the spectra had been performed due mostly to human errors in data collection or for having considered, a posteriori, that some of the analytes used in the mixtures was in bad condition, thus invalidating all of the mixtures that were used. Finally for each one of the families, spectra originating from mixtures of up to four analytes were used.

Results and discussion

The resulting spectra from the previous design were used to validate the system, trying to identify the analytes that contained the previously mentioned mixtures. In Figure 3 examples of identification of some of the mixtures are shown.

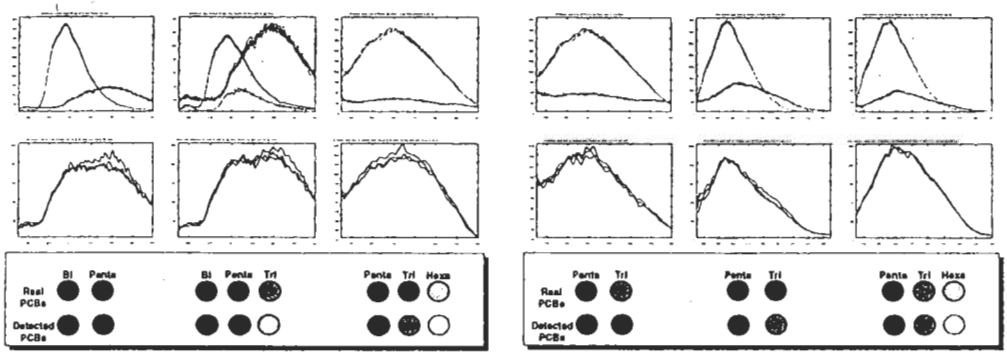


Figure 3: Examples of analyte identification in mixtures generate with different concentration of these analytes using emission, excitation and synchronous spectra

In order to evaluate the efficiency of the system an error function of detected compounds in a mixture was used. This error function is defined by the following expression:

$$E = \frac{NS + BS}{S} \tag{7}$$

Where NS is the number of compounds not detected, BS is the number of compounds badly detected and S is the number of compounds in mixture.

The following table shows averages of the mixture errors taking into account whether or not RBFNs were used and the use of conventional emission, excitation or synchronous spectra:

		Spectra		
		Emission	**Excitation**	**Synchronous**
PCDFs	**Without RBFNs**	0.1286	-	0.0502
	With RBFNs	0.0467	-	0.0000
PCBs	**Without RBFNs**	0.1917	0.2822	0.2615

Table 3: Averages of mixture errors

As can be seen with the PCDF family, both in the use of synchronous spectra as well as in the use of RBFNs, the efficiency of system identification improved noticeably in the preprocessing, reaching to the correct identification of 100% of the analytes. In addition, the advantages of the use of RBFNs are such that better results are obtained with spectra in emission than by using synchronous spectra without RBFNs.

However, in the PCB family the situation varies noticeably. The tests carried out making use of the preprocessing based on RBFNs produced poor results. For this reason the use of this system was ruled out. There were not any improvements using synchronous spectra over other types of spectra. On the contrary, the best results were achieved making use of the emission spectra. The differences in behaviour between families could be a result of the fact that the similarity between PCB spectra is higher than between PCDF spectra, as can be seen in Figure 4. In addition, the Gaussian approximations on which the preprocessing is based with use of RBFNs are not as precise between PCBs as they are between PCDFs

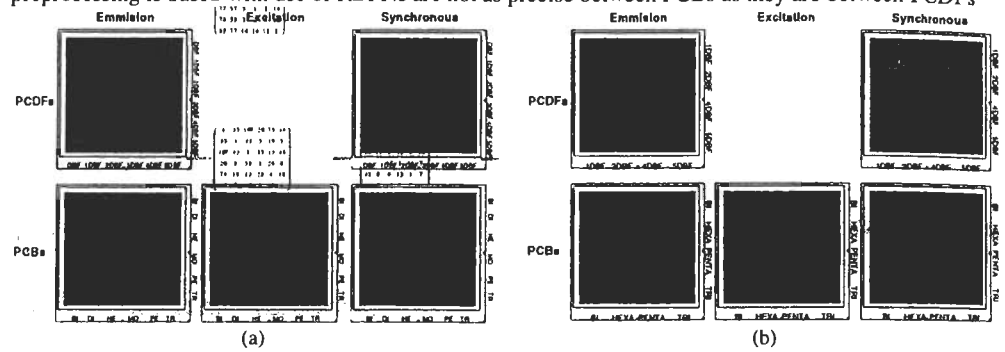


Figure 4: Matrices of correlation between the analytes from the different types of spectra of the PCDF and PCB families a) using real spectra and b) using amplitude vectors a of theirs Gaussian approximations

To illustrate the influence of other aspects in the identification of compounds we present different graphs, with attention to the number of analytes that the mixture to identify has, the concentration of the analytes in the mixture and the influence of the analytes present in the mixtures, Figure 5.

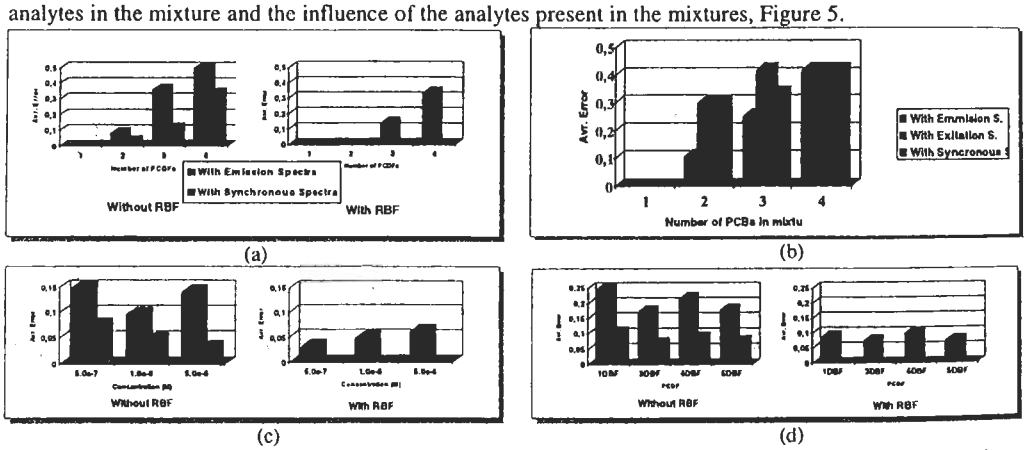


Figure 5: Influence of the number of analytes from the mixture in the average of the errors from the mixtures of a) PCDFs and b) PCBs. Influence of c) the concentration of the analytes of the d) type of analytes present in the mixture over the average of the errors of the mixtures of PCDFs

One can observe that, as expected, the number of analytes that the mixture has influences on the error average in their identification, the identification of mixtures with many analytes being more difficult. In the rest of the characteristics analysed no significant variations are detected.

Conclusions

A computational neural system based in an original hierarchical unsupervised modular adaptive neural network (HUMANN) is presented (HUMANN-CNS). This system consists of two stages: pre-processing and processing module. The processing module is HUMANN which is capable of carrying out blind clustering. Its principal structural characteristic is its modularity, combining several types of neural structures with different types of unsupervised learning.

HUMANN may produce networks close to the architecture of the brain incorporating several new neural mechanisms with biological plausibility. These new mechanisms are the responsible of adaptive behaviour in HUMANN, they are: a) Neural Elimination and b) Neural Generation. The connection structure of this layer maintains the two types of synaptic connections present in the biological neural network, namely the active synapses and the silent synapses.

The preprocessing module consists of a combination of two RBFs and a BPN which allows us to develop a new model for fluorescence spectrum using Gaussian distributions.

We have demostrate that the HUMANN-CNS is a very appropriate method for facing the extremely restricted scope of application of fluorescence spectrometry in the analysis of complex mixtures. This can be seen in the real-world problem solved by HUMANN-CNS in this paper. It has been successfully tested, showing its high efficacy and fitting, in the identification of compounds of environmental interest likes the PCDFs and PCBs without prior knowledge of which (and how many) of these possible compounds can be found in the analysed sample. It has been demonstrate the goodness of our system in complex mixtures of up to four different PCDFs per mixture. We also have introduced an experiment design which optimizes the training set to use.

The obtained results are an important contribution in the environmental analytical chemistry field. This importance is based on that this neural computational method uses only spectral fluorescence data, is very simple, fast and economic method for monitoring of the environment. Another important advantage of our proposal is the possibility to design an on line intelligent environmental monitoring system.

References

1. P. García Báez, C.P. Suárez Araujo, P. Fernández López (2001) Extensión of HUMANN for Dealing with Noise and with Classes of Diferent Shape and Size: A Parametric Study. *Proceedings 6th International Work-Conference on Artificial and Natural Neural Networks (IWANN)* 2, 96-103.
2. P. García Báez, P. Fernández López, C.P. Suárez Araujo. A Parametric Study of HUMANN in relation to the Noise. Application to the Identification of Compounds of Environmental Interest. *System Analysis Modelling Simulation*, in press.
3. O. Hutzinger, S. Safe, V. Zitko (1974) *The Chemistry of PCBs*, CRC Press, Cleveland, OH, 249-251.
4. C. L. Fletcher, W. A. McKay (1993) *Chemosphere* **26** (6), 1041.
5. R.A. Femia, S. Scypinski, L.J. Cline Love (1985) *Environ.Sci.Technol* **19**, 155.
6. I.M. Khasawneh, J.D. Winefordner (1988) *Talanta*, **35** (4), 267.
7. Q. Li, X. Yao, X. Chen, M. Liu, R. Zhang, X. Zhang, Z. Hu (2000) *Analyst* **125**, 2049-2053.
8. J.B.F. Lloyd, I.W. Evett (1977) Prediction of Peak Wavelenghts and Intensities in Synchronously Excited Fluorescence Emission Spectra. *Analytical Chemistry* **49** (12), 1710-1715.
9. S. E. Cabaniss (1991) *Analytical Chemistry* **63**, 1323-1327.
10. W. Lawton, M Martin (1985) The Advance Mixture Problem-Priniples Algorithms. Technical Report IOM384, Jet Propulsion Laboratory.
11. M.T. Musavi, W. Ahmed, K.H. Chan, K.B. Faris, D.M. Hummels (1992) On the Training of Radial Basis Function Classifiers. *Neural Networks* **5**, 595-603.
12. P. Bachiller, R.M. Pérez, P. Martínez, P.L. Aguilar, A.I. Merchán (1998) Optimize on the Size of a Backpropagation NN Solving the Mixture Problem. *Proc. Int. ICSC Symp. on Eng. Int. Syst.*.
13. H.L. Atwood, J.M. Wojtowicz (1999) Silent Synapses in Neural Plasticity: Current Evidence. *Learning & Memory* **6**, 542-571.
14. C. Scharff, J.R. Kirn, M. Grossman, J.D. Macklis, F. Nottebohm (2000) Targeted neuronal death affects neuronal replacement and vocal behavior in adult songbirds. *Neuron* **25** (2), 481-492.