ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

High dimensionality voltammetric biosensor data processed with artificial neural networks

Andreu González-Calabuig, Georgina Faura and Manel del Valle*

Sensors and Biosensors Group, Department of Chemistry, Universitat Autònoma de Barcelona, Edifici Cn, 08193 Bellaterra, Barcelona, Spain

Abstract. This work report the coupling of an array of voltammetric sensors with artificial neural networks (ANN), usually named Electronic Tongue, for the simultaneous quantification of tryptophan, tyrosine and cysteine aminoacids. The obtained signals were compressed using fast Fourier transform (FFT) and then the ANN model was constructed from a set of low-frequency components. An ANN predictive model was obtained by back-propagation, which had 160 input neurons, one hidden layer with 7 neurons and used purelin and satlins functions in the hidden and output layer respectively, trained with a factorial design scheme . The model attained a total normalized root mean square error of 0.032 for an independent test set of data (n=15).

1 Introduction

In the late 90s a new trend started in the field of electrochemical sensors: the use of multiple sensing devices, coupled with advanced statistical tools to solve different types of applications. These sensor arrays were employed to overcome one disadvantage of electrochemical signals: the resolution of highly complex samples in presence of interferents [1]. Since then, a wide variety of data processing techniques have been employed such as principal component analysis (PCA), partial least squares (PLS), principal components regression (PCR) or artificial neural networks (ANNs) [2-5]. A quick evaluation of the concept is that computer science was applied in order to improve the final performance obtained with existing (bio)sensors.

This approach is known as electronic tongue (ET) [6], due to its similarities to the biological taste sense; it has been widely employed in the sensing field in qualitative and quantitative applications such as the determination of the polyphenolic content in wines, the prediction of the wine score or pollutant monitoring in wastewaters, among many others [7-10]. When sensors used are of the voltammetric type, high dimensionality, normally tri-linear, signals are obtained, which poses difficulties in obtaining the response models.

The study case presented here is the coupling of cyclic voltammetry (CV) responses obtained from an amperometric biosensor electrode array, compressed with Fast Fourier Transform (FFT) and processed with ANNs to build a predictive model. The application developed is the quantitative prediction of the contents of tryptophan, tyrosine and cysteine aminoacids in a certain sample. These aminoacids are chosen because they are the ones that are oxidizable compounds, i.e. the ones that can be determined with the voltammetric technique.

* E-mail: manel.delvalle@uab.cat

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

2 Case of study/description of the data set

In this case the species that are quantified are the three aminoacids: tryptophan (TRP), tyrosine (TYR) and cysteine (CYS); these aminoacids have similar voltammetric signals. The system array was formed by five sensors; each sensor was a voltammetric epoxy-graphite electrode, bulk-modified with different catalysts (metal or metal-oxide nanoparticles) in order to induce slightly differentiated responses for each sensor. The CV signal from each of these sensors was a vector of 556 current intensities (at each probed polarization potential); thus the departure information recorded per each sample was a matrix of 556 x 5 numeric values (Figure 1A).

The training set was designed from a full factorial design (3 levels x 3 factors, the three aminoacids), with concentrations ranging from 0 to 30 mM; an external test subset was also measured, formed by 15 samples with their concentrations randomly distributed inside the factorial design, as can be seen in Figure 1B.

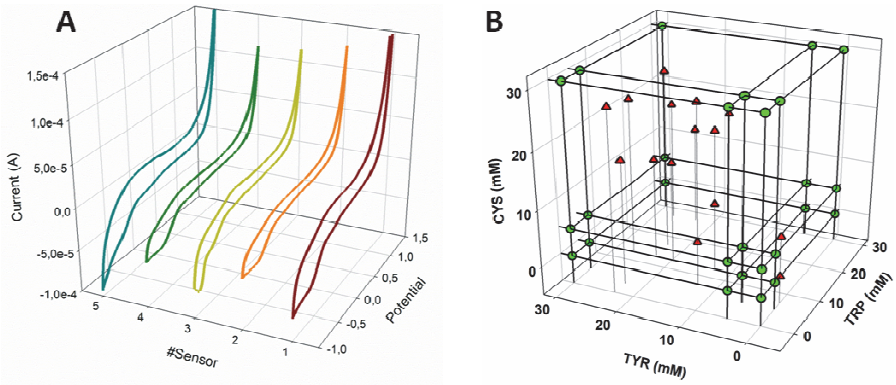

Fig. 1: (A) Representation of the obtained CV signal per each sample. (B) Experimental design employed, including training (●) and test (▲) subsets.

3 Data preprocessing

ANNs are powerful modeling tools but have some drawbacks, making them less used in the field of chemometrics than other procedures such as PLS or PCR. Some of the drawbacks are the time required to build the model or the possibility of *overfitting* the model if high dimensionality data are used. Raw data is often pre-processed in order to reduce its complexity while preserving the relevant information; thus these drawbacks may be avoided.

3.1 Data scaling

In electrochemical applications it is important to correct any potential drift of the electrochemical sensors; if responses change with time, the prediction capabilities of the model may be degraded. A first approach is to correct the signals obtained from the sensor array. In this case, previous to any compression, we chose to scale the responses of each sensor, as it potential effect on the ANN model was unclear.

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

3.2 **Data compression**

Data was next compressed evaluating two different procedures, by means of fast Fourier transform (FFT) or through discrete Wavelet transform (DWT). The signal of each sensor was compressed individually, to avoid losses in the relevant information, and the resulting coefficients were unfolded and recovered as a single vector. For a given CV signal, the goodness of the compression step was evaluated following reported methods [11, 12]; therefore, by comparing the original signal (a_i) with the reconstructed signal (b_i) after the compression step and calculating the correlation (R^2) or the *fc* factor, defined as the ratio between the areas covered by the two curve signals (equation 1). This allowed us to determine the performance of the compression step in terms of information loss.

$$fc = \frac{\sum_i (max < a_i, b_i > - |a_i - b_i|)}{\sum_i (min < a_i, b_i > - |a_i - b_i|)} \qquad (1)$$

Taking into account the options described the variants tested were the following: compression to 16 or 32 Fourier coefficients, compression using *Daubechies 3* mother function through levels 4, 5 or 6. After evaluating the performance of the different compression procedures, represented in Figure 2, the chosen method was fast Fourier transform with 32 coefficients, as this kind of compression performs especially well in compression of CV signals, such is this case.

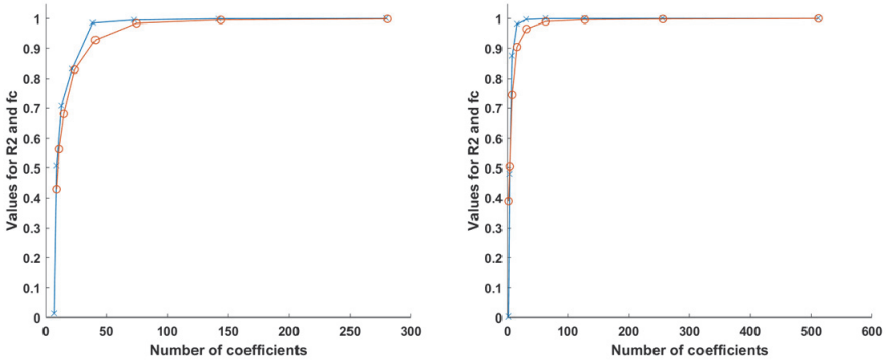

Fig. 2: Representation of the coefficient of determination (R^2), blue star, and *fc*, red circle, against the number of coefficients used from the raw voltamograms with the reconstructed signal for FFT (rigth) and DWT (left).

4 Artificial Neural Networks model building

Once the different subset data was collected, voltammograms were compressed by means of FFT, so they suited to feed the different ANN models. For simplicity, the ANNs evaluated were composed by one input layer, one hidden layer and one output layer; more complex models were discarded given their complexity, and that a correct behavior of similar designs with related cases was obtained previously [10,11].

The first configuration detail considered was choosing the appropriate ANN architecture; this is usually done by trial-and-error procedure due to the difficulties to

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

predict the best model in advance as several parameters may show their effect (compression pretreatment, number of neurons in the hidden layer, transfer functions used, etc.) [13]. The parameters considered to evaluate performance were: normalized root mean square error (NRMSE, Eq. 2), slope, intercept and correlation of the obtained vs. expected concentrations of the comparison graphs. The last three (x_j) were treated globally as deviations from their ideal value for easier comparative purposes (Eq. 3 for correlation and slope, and Eq. 4 for intercept). Figure 3 shows the plots comparing the models evaluated when using a total of 160 Fourier coefficients.

$$\mathrm{NRMSE} = \sum_i \frac{\mathrm{RMSE}_i}{c_{i,max} - c_{i,min}} \quad (2)$$

$$\Delta x = \sum_j |1 - x_j| \quad (3)$$

$$\Delta x = \sum_j |x_j| \quad (4)$$

As previously commented, the samples from the training subset were used to build the ANN model while performance of the model was evaluated from the comparison graphs of predicted aminoacid concentrations in the test subset samples. As the test subset is an external set that has not been used in the modeling; the goodness of fit for this subset is a good indicator evaluating modeling performance.

After the evaluation of the different topologies, the final ANN architecture had 160 neurons (5 sensors × 32 FFT coeffs.) in the input layer, 7 neurons and *purelin* transfer function in the hidden layer and 3 neurons and *satlins* transfer function in the output layer, providing simultaneously the concentration of the three compounds considered. Total number of ANN architecture configurations evaluated was 192.

The correct prediction ability of the model could be visualized in the predicted vs. expected concentration comparison graphs, for training and testing subsets, for each of the three aminoacids considered, as shown on the Figure 4. The detailed parameters of the three comparison graphs are summarized in Table 1.

SUBSET	AA	Correlation	Slope	Intercept (10^{-4}M)	NRMSE
TRAIN	TRP	0.999	0.924±0.004	0.105±0.007	0.018
	TYR	0.989	0.872±0.010	0.176±0.017	
	CYS	0.998	0.973±0.012	0.046±0.008	
TEST	TRP	0.960	0.853±0.035	0.315±0.054	0.032
	TYR	0.952	0.807±0.036	0.363±0.070	
	CYS	0.967	0.920±0.068	0.053±0.118	

Table 1: Results of the fitted regression lines for the comparison between obtained vs. expected values, both for the training and testing subsets of samples and the three considered species (intervals calculated at the 95% confidence level).

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

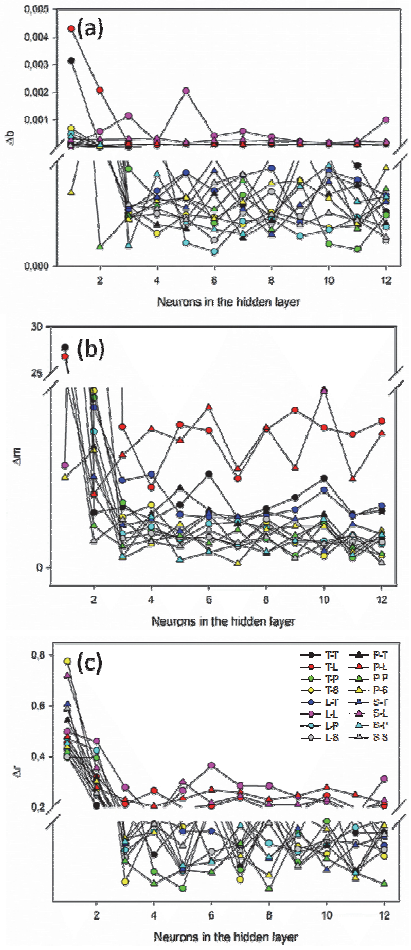

Fig. 3: Detailed results of the ANN optimization for 160 FFT coefficients, non-centred data and a residual error of 10^{-10} M. Obtained (a) correlation coeffs, (b) slopes, (c) intercepts values from the obtained *vs.* expected comparison graphs for the testing subset are plotted against the number of neurons in the hidden layer when employing different transfer functions (L: *logsig*, P: *purelin*, S: *satlins*, T: *tansig*).

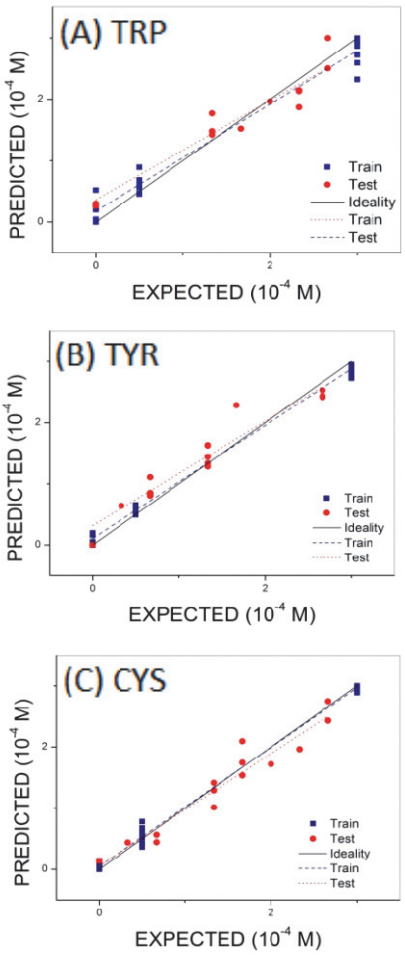

Fig. 4 Comparison of the expected vs. the predicted concentrations for the optimized FFT-ANN model, for the three aminoacids studied.

A satisfactory trend is obtained for both subsets, with regression lines values very close to the theoretical ones, slopes and intercepts equal to 1 and 0 respectively. Nevertheless, the training subset have a better correlation coefficients ($r \geq 0.99$) than

the test subset ($r \geq 0.95$) but it is expected as the train subset is used to optimize the architecture, therefore the model is tailored to fit this data while the test subset is not used at all. The detailed results show promising results for the test subset as the NRSME for the three compounds is 0.032.

5 Conclusions

This work combines the use of an array of electrochemical sensors and advanced mathematical tools such as artificial neural networks have proved to be useful to resolve ternary mixtures of aminoacids. The reduction of the data dimensionality, by means of fast Fourier transform (or discrete wavelet transform), allows a considerable reduction in the training time and avoids the *overfitting* of the different ANN models.

Acknowledgements

Financial support for this work was provided by the Spanish Ministry of Economy and Innovation, MINECO (Madrid) through project CTQ2013-41577-P. Andreu González-Calabuig thanks Universitat Autònoma de Barcelona for the PIF fellowship. Manel del Valle thanks the support from program ICREA Academia.

6 References

[1] Vlasov Y, Legin A, Rudnitskaya A, Di Natale C, D'amico A. Nonspecific sensor arrays (" electronic tongue") for chemical analysis of liquids. *Pure and Applied Chemistry*. 77(11):1965-83, International Union of Pure and Applied Chemistry, 2005.

[2] Jolliffe I. *Principal Component Analysis.* Encyclopedia of Statistics in Behavioral Science: John Wiley & Sons, Chichester, 2005.

[3] Wold H. *Partial Least Squares*. Encyclopedia of Statistical Sciences, Wiley, Chichester, 2006.

[4] Jolliffe IT. A note on the use of principal components in regression. *Applied Statistics*, 31:300-303, Royal Statistical Society, 1982.

[5] Hanrahan G. Computational neural networks driving complex analytical problem solving. *Analytical Chemistry*.82(11):4307-4313, American Chemical Society, 2010

[6] del Valle M. Sensor arrays and electronic tongue systems. *International Journal of Electrochemistry*. 2012; Hindawi, 2012.

[7] Ni Y, Kokot S. Does chemometrics enhance the performance of electroanalysis? *Analytica Chimica Acta,* 626(2):130-146, Elsevier, 2008.

[8] Cetó X, Gutiérrez JM, Gutiérrez M, Céspedes F, Capdevila J, Mínguez S, et al. Determination of total polyphenol index in wines employing a voltammetric electronic tongue. *Analytica Chimica Acta*, 732:172-179, Elsevier, 2012

[9] Cetó X, González-Calabuig A, Capdevila J, Puig-Pujol A, del Valle M. Instrumental measurement of wine sensory descriptors using a voltammetric electronic tongue. *Sensors Actuators B: Chemical*, 207:1053-1059, Elsevier, 2015.

[10] Cetó X, González-Calabuig A, del Valle M. use of a bioelectronic tongue for the monitoring of the photodegradation of phenolic compounds. *Electroanalysis*, 27(1):225-233, Wiley, 2015.

[11] Moreno-Barón L, Cartas R, Merkoçi A, Alegret S, del Valle M, Leija L, et al. Application of the Wavelet transform coupled with artificial neural networks for quantification purposes in a voltammetric electronic tongue. *Sensors Actuators B: Chemical*, 113(1):487-499, Elsevier, 2006;

[12] Moreno-Barón L, Cartas R, Merkoçi A, Alegret S, Gutiérrez JM, Leija L, et al. Data compression for a voltammetric electronic tongue modelled with artificial neural networks. *Analytical Letters,* 38(13):2189-2206, Taylor & Francis, 2005.

[13] Despagne F, Massart DL. Neural networks in multivariate calibration. *Analyst*, 123(11):157R-178R, Royal Society of Chemistry, 1998.