Quantitation in Multianalyte Overlapping Peaks from Capillary Electrophoresis Runs Using Artificial Neural Networks

S. Sentellas, J. Saurina*, S. Hernández-Cassou, M.T. Galceran, and L. Puignou

Department of Analytical Chemistry, University of Barcelona, Diagonal 647, E-08028, Barcelona, Spain

Abstract

The potentiality of artificial neural networks for multicomponent analysis in unresolved peaks from capillary electrophoresis (CE) is evaluated. The system chosen consists of mixtures of three ebrotidine metabolites, which cannot be successfully separated by CE. Data selected for analysis consist of UV spectra taken at the maximum of the CE peak. The most dissimilar analyte, in terms of spectral differences, is accurately quantitated in any type of mixture with an overall prediction error of 5%. Because of the strong interference of the two most overlapped compounds, a preliminary procedure for spectral data filtering based on principal component analysis is performed to improve their quantitation.

Introduction

The application of mathematical tools to separation techniques has been specially developed for liquid chromatography. Examples describing procedures for resolving overlapping peaks (1,2), checking the peak purity (3), or improving the quantitation (4,5) have been reported. However, the linkage of chemometrics to capillary electrophoresis (CE) has been scarcely addressed. For instance, the optimization of CE separations using experimental design (6) and multiple linear regression (7), the quantitation with partial least-square regression (8), or peak resolutions with alternating least-squares (9) have been described.

An artificial neural network (ANN) is a data processing system that is based on neuron relationships occurring in the human brain. ANN architectures consist of highly interconnected node structures arranged in layers. The pioneering developments in ANN date back to the 1940s (10,11). Although the simplicity of this approach made it a very promising tool for data processing, ANN fell into disuse for several decades because of the need for computer facilities. Recent developments in the computational and mathematical fields have contributed to increasing interest in these chemometric techniques. Today, ANNs are being used extensively for solving modeling problems in analytical chemistry (12-14).

The application of ANN to electrophoretic techniques has been seldom described to date, although this field is receiving increasing attention in recent years. A few examples dealing with ANN in combination with CE have been reported in the literature. ANN has mainly been used in experimental design for optimizing separation and improving the quantitation in poorly resolved peaks. For instance, back propagation (BP)- ANN has been used to optimize the separation by capillary zone electrophoresis (CZE) (15–17). Migration studies for the improvement of the resolution have also been reported in micellar electrokinetic capillary chromatography by using BP-ANN (18). Other examples deal with ANN classifications of electrophoretic profiles for citrus juice characterization (19), tumor diagnosis (20), or pentosan polysulfate characterization (21). The quantitative determination of amino acids in complex overlapping peaks from the analysis of either spectral or electrophoretic profiles has been addressed by Latorre et al. (22).

The application of ANN to CE analysis creates special difficulties that are not encountered in other separation techniques, such as high-performance liquid chromatography (HPLC). This concerns some particularities of the technique (e.g., temperature gradients associated to the Joule heating created, comigrations with the electroosmotic flow, and stacking effects caused by matrix composition). As a result, specific CE shortcomings (such as variabilities in migration time, peak broadening and resolution, baseline drifts, etc.) may arise.

The present paper is aimed at exploring the performance of ANN for improving the quantitation of analytes in unresolved CE peaks. Here, some of the drawbacks commented on earlier have been corrected with a suitable data pretreatment, yet others have been overcome by proper ANN training. The system chosen for the study consists of mixtures of three ebrotidine metabolites currently present in human urine, namely 4-bromobenzensulfonamide (compound A), *N*-(2-methylsulfonyl-ethylamin-methylen)-4-bromobenzensulfonamide (compound C). The significance of ebrotidine is justified

^{*} Author to whom correspondence should be addressed: email xavi@apolo.qui.ub.es.

because it is used as an antisecretory drug in the treatment of gastric ulcers (23). The analytical methods for the determination of ebrotidine and its metabolites in corporeal fluids are based on HPLC and CZE techniques with UV–vis (24–26) and mass spectrometry (26–28) detection.

Compounds A, B, and C are poorly ionized under the experimental conditions for carrying out the CZE separation of common ebrotidine metabolites. As a result, they are hard to separate by this technique and, thus, comigrate. ANN can take advantage of the richer information contained in the multidimensional CE–diode array spectrophotometric (DAS) data, which allow such compounds to be discriminated and quantitated in unresolved CE peaks. In particular, spectra taken at peak maximum have been chosen for ANN analysis.

Theory

General strategy

In this section, some general guidelines concerning the application of ANN to CE data and a theoretical background of ANN are discussed. The resolution of analytical problems dealing with the simultaneous determination of various (or numerous) compounds commonly involves the application of separation techniques. Among them, CE is gaining popularity in the analytical field because of its great possibilities.

The first step in the development of a new CE method or application is the optimization of the separation conditions in order to achieve a full resolution of all analytes of interest in the samples. The electrophoretic separation of analytes currently provides sufficient selectivity for their determination using conventional univariate calibration. As a result, the use of more sophisticated chemometric algorithms for improving the quantitation is not necessary. However, despite the great separation performance of CE, a full resolution of all analytes might not be achieved experimentally. Under these circumstances, the application of chemometrics, particularly ANN, may be especially recommendable to improve the accuracy of the determination.

The application of ANN to poorly resolved peaks requires the achievement of various conditions that the analyst should keep in mind.

Reproducibility

In some cases, the variability of the migration time affects the reproducibility of the CE data. Spectral information is more robust in front of this factor, whereas electrophoretic profiles may be severely influenced by this variation. In these circumstances, data pretreatments for peak shifting correction may still be used to correct this variability.

Dissimilarity

Two or more comigrating analytes could be resolved chemometrically when they present a reasonable degree of dissimilarity in their respective responses. Although full selectivity is not required, each analyte obviously has to be distinguishable from each other. Therefore, the particular features in the response of the analyte may be advantageously used for its determination in the presence of other species. Otherwise, when the similarity values of the analyte profiles are extreme, the systems may not be solved chemometrically, thus, the strategy for the resolution of the analytical problem should be reconsidered. In CE with a fast scanning detection system [e.g., CE-diode array detection (DAD)], spectrometric and electrophoretic profiles can be used as analytical data. Therefore, the dissimilarity of the profiles of the analytes can be evaluated in both spectrometric and time domains in order to ascertain the type of data that is more suitable for further chemometric analysis. Experimentally, dissimilarities are checked from the profiles obtained from pure standards of each component by calculating the correlation coefficient between the corresponding profiles.

ANNs

ANNs consist of highly interconnected node structures arranged in layers. The most common ANN learning process used in analytical chemistry is the BP method. A more detailed description is given elsewhere (12–14).

The performance of ANN for predicting the analytes was evaluated by leave-one-out cross validation. In this approach, each sample is predicted using the remaining samples as the training set. In addition, each analyte was quantitated individually from an ANN trained specifically for it, as suggested by Courtois et al. (14). This was because a simpler ANN could provide better models for predicting each particular analyte than more complex architectures.

The architecture of an ANN in these studies consisted of a three-layer structure. The input layer contained the electrophoretic spectral response, the output layer delivered the concentration of a particular analyte, and the hidden layer correlated the information between inputs and outputs. More complex cases can be designed in which two or more hidden layers are included.

The activation of a hidden neuron is defined as the sum of weighed inputs to its node:

$$n_k = \sum_j w_{jk} \bullet x_j + b_k$$
 Eq. 1

where n_k is the activation value of neuron k, b_k is the bias value, x_j is the input variable for a given sample, and w_{jk} is the weight. The relationship between n_k and the values given by the neuron (a_k) was established with a sigmoid transfer function, as follows:

$$a_k = \sum_{1}^{k} \frac{1}{1 + e^{-n}}$$
 Eq. 2

The analyte concentration (y) is correlated with these transferred values using the equation:

$$y = \sum_{1}^{k} a_k \bullet w_k + b$$
 Eq. 3

where y is the concentration predicted and w_k and b are the corresponding weight and bias values for the hidden nodes of k. The BP network "learns" by adjusting its weights and biases iteratively from random values given initially, until the average error reaches a desired value (here, the value was fixed at 0.05%) or the number of iterations exceeds a previously fixed maximum (here, 50,000 epochs).

The training or calibration step involved the use of proper standards to establish the mathematical relationship between the experimental response and the analyte concentration. Although there are no specific rules, the number of standards should increase with the complexity of the system (number of analytes, degree of overlapping, etc.). In general, the number of standards should be higher than the number of weights and biases to be fitted in the model. Furthermore, standards should cover the variability displayed by the unknown samples in terms of concentration, matrix composition, etc.

Once the training process was completed, the optimized feedforward network architecture could be used for predicting the output values for unknown samples from the corresponding input values.

Experimental

Chemicals

4-bromobenzensulfonamide (compound A), *N*-(2-methylsulfonyl-ethylamine-methylene)-4-bromobenzensulfonamide (compound B), and *N*-(2-methylsulfinyl-ethylamin-methylene)-4-bromobenzensulfonamide (compound C) were purchased from Grupo Ferrer (Barcelona, Spain). Work solutions were prepared in acetonitrile (Merck, Darmstadt, Germany). Reagents used for the preparation of carrier electrolyte solution were acetic acid and ammonia (Merck).

Samples

The single-, two-, and three-component samples were prepared according to full factorial designs at two concentration levels. The uppercase letter indicates the metabolite, and the number refers to the concentration level. The low level that contained 100 µg/mL of analyte is referred to as 1; the high level that contained 200 µg/mL is referred to as 2. For instance, A1B2 is the two-component mixture that consisted of 100 µg/mL of compound A and 200 µg/mL of compound B.

Instrumentation

CE runs were carried out with a Beckman (Fullerton, CA) P/ACE System 5500 system with a DAS detector. A fused-silica capillary (Supelco, Bellefonte, PA) of 50–57 cm and 75-µm i.d. was used. Spectra were acquired at regular steps of 0.26 s during the electropherogram. Data were converted into ASCII files for mathematical treatment with a PC using Beckman P/ACE station software (version 1.0). Eleven working wavelengths (220–280 nm) were chosen for analysis.

CE conditions

A 50mM acetic–acetate buffer at pH 5.7 (adjusted with ammonia) was used as the carrier electrolyte. CE potential was set at 20 kV, and the temperature was held at 25°C. The sample was hydrodynamically injected (0.5 psi) for 4 s.

The capillary was pretreated by rinsing (20 psi) 1M sodium hydroxide solution for 5 min followed by a 5-min ultrapure water rinse. It was then stabilized with the carrier electrolyte for 1 h. The capillary was rinsed with the buffer for 2 min before each run.

Data pretreatment

A data pretreatment was performed in order to correct peak shifting and spectral drifts. The variability of the migration time of the CZE peaks was corrected with a peak alignment procedure. A peak maximum at 220 nm was used to position the electropherogram. The spectrum at the beginning of the time window was subtracted over the whole time range in order to correct spectral drift. Refractive index correction was carried out by subtracting the absorbances at 290 nm from the time profile at each wavelength. A more detailed description is given elsewhere (9).

Software

Matlab (Natick, MA) for windows (version 4.2) was used for shift adjustment. The ANN software was Simtelnet (Cheshire, U.K.) EASYNN60 program (available from ftp://ftp.rediris.es/mirror/ Simtelnet/win95/neural/00_index.txt).

Results and Discussion

Figure 1 shows the spectra of compounds A, B, and C obtained from pure standards. It can be seen that compound A displays the most distinguishable spectrum, whereas spectra of compounds B and C are highly correlated. These spectral features can be related to the molecular structures of these compounds; compounds B and C are almost identical (they only differ in the presence of a sulfoxide or sulfone group), and in contrast, compound A has a rather different chemical structure. Owing to the lack of selectivity for the compounds under study, their determination should be based on the richer information contained in the multivariate spectral data.

Data pretreatment

The high voltage applied during the CE run generates a heating by Joule effect that should be dissipated appropriately. Heating is problematic because it can cause nonuniform temperature gradients and subsequent band broadening. Additionally, the run-torun reproducibility of migration time, stability of the baseline, and resolution can be seriously affected by this phenomenon. In this paper, the variability of the migration time and the baseline drift were minimized with a proper data pretreatment (see the Experimental section). Unfortunately, variations in the resolution could not be corrected easily. Furthermore, the EOF peak signal caused by the sample solvent overlapped with peaks of compounds A, B, and C and, thus, interfered with their analysis. This represented an additional drawback for the data treatment as the variability of this peak was especially remarkable. Therefore, from





resolution and EOF changes, two additional sources of data variance should be modeled implicitly during the ANN learning process.





Figure 3. Variation of the training and prediction errors with the number of iteration cycles for compound A. Test sample, A1B2C2. Symbols are given for calibration (\bigcirc) and prediction (\blacklozenge).



ANN data analysis

In the present study, ANN has been used for solving the lack of selectivity, and then for quantitating these analytes in single-, two-, and three-component mixtures. For this purpose, BP-ANN with sigmoid transfer neurons were trained. The corresponding ANN architecture for these cases consisted of a three-layer ANN containing the inputs that corresponded to the absorbance values taken at each working wavelength (alternatively, the scores of principal components, discussed later) of a given sample, various nodes in the hidden layer, and one node as output. In all cases, the values of momentum and learning rate were 0.3 and 0.6, respectively. These parameters were used for optimizing the weight changes in each iteration cycle.

The number of nodes to be used in the hidden layer was optimized by minimizing the prediction error. In a majority of cases, three nodes were found to be optimal. The influence of the number of nodes on the predictions is illustrated in Figure 2, in which the concentration of compound A in the three-component mixture A1B2C2 was estimated as a function of this parameter. As shown in this figure, poor predictions were obtained when using one or two nodes, although the accuracy was acceptable for more than three nodes. In addition, the performance of the quantitation was not affected significantly when increasing the number of nodes. Although good estimations were obtained with more than three nodes, these more complex architectures could lead to overfitting and may provide wrong predictions.

Data autoscaling and normalization were two preprocessing procedures that were evaluated in order to ascertain their effect on the predictions. Results were compared with those obtained from the original data using compound A as a model. The overall prediction errors were similar for original, autoscaled, and normalized data. As a conclusion, these preprocessing procedures did not significantly improve the quantitation, therefore nonpreprocessed data were used in further studies.

The stability of the ANN structure for training and prediction was checked in a series of 50 runs, each one using initial random estimates of weights and biases. A1B2C2 was chosen as a test sample, which was predicted using the remaining samples as standards. As a result, the average concentration value obtained in this series of calculations was 96 mg/L (actual value 100 mg/L), and the relative standard deviation was 4.1%. These values sug-

gested that the net training was very efficient at providing similar predictions for any run initialized randomly.

The variation of the error in the training (calibration) and prediction steps versus the number of iterations is shown in Figure 3 for compound A. As in the previous study, A1B2C2 was selected as the test sample. In the initial iteration cycles, the training error decreased dramatically with the number of epochs, yet from 1000 cycles the improvement in the fitting of the data was more gradual. However, the number of steps required to reach a minimum in the prediction of a test sample was considerably higher. In general, between 20,000 and 50,000 epochs may be necessary to train the net to reach the optimum results. The conclusion of this study can be extended to compounds B and C and to the other test samples.

The quantitation of compound A in a series of single-, two-, and three-component samples is detailed in Figure 4. The bars represent the average concentration predicted, and the dots indicate the corresponding actual values. The precision of these predictions was evaluated from three independent training processes after the corresponding random initializations of network biases and weights. The precision range is also plotted. The concordance between the actual and predicted values is considered acceptable, with an overall prediction error of 5.0%. These results proved the possibility of using these modeling techniques for the quantitation of a given nonresolved compound in a complex mixture when the analyte spectrum differs sufficiently from the others.

The quantitation of compounds B and C using the same modeling strategy was deficient because of their high spectral overlapping. This finding suggested the limitations of ANN for making accurate predictions for strongly correlated data. However, the total amount of these compounds (i.e., B + C concentrations) was estimated with reasonable concordance, with the overall prediction error at 11.9%. Although this total concentration did not provide information about each species, it served as biochemical estimation of the extension of some overall metabolism pathways. Indeed, the chemical and biochemical behavior of the sulfoxide and sulfone metabolites are similar.

In order to try to improve the determination of these compounds separately, a preliminary data treatment for removing correlated information was checked. This strategy consisted of data filtering using principal component analysis (PCA). The original information contained in the 11 experimental variables was concentrated into 6 principal components, which were needed to keep the significant variance. Predictions for compounds B and C using this type of data improved notably with respect to the previous approach, with overall errors at approximately 10%. These results were considered acceptable when taking into account their strong spectral overlapping.

In order to check the performance of ANN with respect to other chemometric methods, results from this analysis were compared with those obtained with various multivariate calibration methods [e.g., principal component regression (PCR), linear and nonlinear partial least-square regression (PLS), multivariate curve resolution based on alternating least-squares (ALS), etc.]. In the case of compound A, prediction errors were 9.0, 7.6, and 12.0% for PCR, PLS, and ALS, respectively. For compounds B and C, the error ranged between 9 and 20%. From these values, ANN provided, at least, results as accurate as any other chemometric method. Additionally, ANNs were especially appropriate to detect the absence of the analyte (i.e., a concentration of 0 mg/L) or predict low concentrations, which was a more difficult task when using the other chemometric methods.

Conclusion

This paper is focused on the application of ANN to the evaluation of CZE data. The general strategy for quantitating in overlapping peaks is discussed. The application of chemometrics, par-

ticularly ANN, to CE is still in its infancy in comparison with other separation techniques, so there is a lot of work to do in this field. Similar to HPLC and other separation techniques, CE peaks can be analyzed using ANN, although additional shortcomings related to the complexity of the technique and the generated CE data have to be solved. The ANN performance was checked in various examples of different complexity in terms of spectral correlation. The quantitation of highly similar compounds was severely interfered, but the more dissimilar could be determined accurately. A potential way of improving the determination of highly correlated compounds was based on the application of PCA filtering in order to remove redundant information. Under this approach, results improved significantly and even highly correlated compounds could be predicted with reasonable accuracy. In addition, the general conclusions drawn with respect to the applicability of ANN to electrophoretic data for improving the quantitation can be extended to many other examples in which a full separation is not achieved.

Acknowledgments

This study was partially financed by projects BQU2000-0788 and SGR99-00049. Sònia Sentellas thanks the University of Barcelona for a grant (Beca de Formació en la Recerca i la Docència). The authors are obliged to Grupo Ferrer Internacional for providing ebrotidine and its metabolites.

References

- R. Tauler, S. Lacorte, and D. Barceló. Application of multivariate selfmodeling curve resolution to the quantitation of trace levels of organophosphorus pesticides in natural waters from interlaboratory studies. J. Chromatogr. A 730: 177–83 (1996).
- R. Tauler and D. Barceló. Multivariate curve resolution applied to liquid-chromatography diode-array detection. *Trends Anal. Chem.* 12: 319–27 (1993).
- A. de Juan, S.C. Rutan, R. Tauler, and D.L. Massart. Comparison between the direct trilinear decomposition and the multivariate curve resolution-alternating least-squares methods for the resolution of 3-way data sets. *Chemom. Intell. Lab. Sys.* **40**: 19–32 (1998).
- M.D. Gil García, A. Garrido Frenich, J.L. Martínez Vidal, M. Martínez Galera, A. Muñoz de la Peña, and F. Salinas. Resolution of overlapping peaks in HPLC with diode-array detection by application of partial least-squares calibration to cross-sections of spectrochromatograms. *Anal. Chim. Acta* 348: 177–85 (1997).
- A. Garrido Frenich, M. Martínez Galera, M.D. Gil García, J.L. Martínez Vidal, M. Catasus, L. Martí, and M.D. Mederos. Resolution of HPLC-DAD highly overlapping analytical signals for quantitation of pesticide mixtures in groundwater and soil using multicomponent analysis and neural networks. *J. Liq. Chromatogr. Rel. Tech.* 24: 651–68 (2001).
- K.D. Altria, B.J. Clark, S.D. Filbey, M.A. Kelly, and D.R. Rudd. Application of chemometric experimental designs in capillary electrophoresis—A review. *Electrophoresis* 16: 2143–48 (1995).
- H. Wan, M. Ohman, and L.C. Blomberg. Chemometric modeling of neurotransmitter amino acid separation in normal and reversed migration micellar electrokinetic chromatography. *J. Chromatogr. A* 916: 255–63 (2001).
- 8. R.M. Latorre, J. Saurina, and S. Hernández-Cassou. Determination of

amino acids in overlapped capillary-electrophoresis peaks by means of partial least-squares regression. *J. Chromatogr. A* **871:** 331–40 (2000).

- 9. S. Sentellas, J. Saurina, S. Hernández-Cassou, M.T. Galceran, and L. Puignou. Resolution and quantitation in poorly separated peaks from capillary-zone-electrophoresis using 3-way data-analysis methods. *Anal. Chim. Acta* **431**: 49–58 (2001).
- 10. W.S. McCulloch and W. Pitts. A logical calculus of ideas immanent in nervous activity *Bull. Math. Biophys.* 5: 115–33 (1943).
- 11. D.O. Hebb. *The organization of Behavior*. Wiley, New York, NY, 1949.
- 12. S. Haykin. *Neural Networks*. Macmillan College Publishing, Englewood Cliffs, NJ, 1994.
- J. Župan and J. Gasteiger. Neural Networks for Chemists: an Introduction. VCH, Weinheim, Germany, 1993.
- 14. S. Courtois and R. Phan-Tan-Luu. Neural networks applied to the choice of an optimal experimental design. *Analysis* **26:** 304–10 (1998).
- J.P. Wolbach, D.K. Lloyd, and I.W. Wainer. Approaches to quantitative structure-enantioselectivity relationship modeling of chiral separations using capillary electrophoresis. *J. Chromatogr. A* **914**: 299–314 (2001).
- G. Bocaz-Beneventi, R. Latorre, M. Farkovà, and J. Havel. Artificial neural networks for quantitation in unresolved capillary-electrophoresis peaks. *Anal. Chim. Acta* 452: 47–63 (2002).
- E.F. Hilder, C.W. Klampfl, and P.R. Haddad. Pressurized-flow anionexchange capillary electrochromatography using a polymeric ionexchange stationary phase. *J. Chromatogr. A* 890: 337–45 (2000).
- Q.F. Li, Y.Y. Zhou, H.W. Wang, H.Y. Zhang, S.H. Liu, X.G. Chen, and Z.D. Hu. Application of artificial neural networks in multifactor optimization of selectivity in capillary electrophoresis. *Anal. Lett.* 33: 2333–47 (2000).
- P.F. Cancalon. Analytical monitoring of citrus juices by using capillary electrophoresis. J. AOAC Int. 82: 95–106 (1999).
- 20. R. Zhao, G. Xu, B. Yue, H.M. Liebich, and Y. Zhang. Artificial neural-

network classification based on capillary electrophoresis of urinary nucleosides for the clinical diagnosis of tumors. *J. Chromatogr. A* **828**: 489–96 (1998).

- B. Schirm, H. Benend, and H. Watzig. Improvements in pentosan polysulfate sodium quality assurance using fingerprint electropherograms. *Electrophoresis* 22: 1150–62 (2001).
- Ř. Latorre, S. Hernández-Cassou, and J. Saurina. Artificial neural networks for quantitation in unresolved capillary-electrophoresis peaks. J. Sep. Sci. 24: 427–34 (2001).
- S.J. Kounturek, N. Kwiecien, E. Sito, W. Obtulowicz, K. Kaminski, and J. Oleksy. Effects of ebrotidine on aspirin-induced gastricmucosal damage and blood-flow in humans. *Scand. J. Gastroenterol.* 28: 1047–50 (1993).
- E. Rozman, M.T. Galceran, Ll. Anglada, and C. Albet. Metabolites of ebrotidine, a new H-2-receptor antagonist, in human urine. *J. Pharm. Sci.* 83: 252–54 (1994).
- E. Rozman, M.T. Galceran, Ll. Anglada, and C. Albet. Investigation of the metabolism of ebrotidine in human urine by liquid chromatography-atmospheric pressure chemical-ionization mass spectrometry. *Drug Metab. Dispos.* 23: 976–81 (1995).
- S. Sentellas, L. Puignou, E. Moyano, and M.T. Galceran. Determination of ebrotidine and its metabolites by capillary electrophoresis with UV and mass-spectrometry detection. *J. Chromatogr. A* 888: 281–92 (2000).
- E. Rozman, M.T. Galceran, and C. Albet. Determination of ebrotidine and its metabolites in human urine by reversed-phase ion-pair high-performance liquid chromatography. J. Chromatogr. B 688: 107–15 (1997).
- E. Rozman, M.T. Galceran, and C. Albet. Ebrotidine and its metabolites studied by mass-spectrometry with electrospray-ionizationcomparison of tandem and in-source fragmentation. *Rapid Commun. Mass Spectrom.* 9: 1492–98 (1995).

Manuscript accepted February 4, 2003.