DATA PROCESSING AND PATTERN RECOGNITION IN HIGH-THROUGHPUT CAPILLARY ELECTROPHORESIS

G.erardo A. Ceballos[1], Jose L. Paredes[1], and Luis Hernández[2]

[1]Biomedical Engineering Research Group. Electrical Engineering Department, University of Los Andes, Mérida, Venezuela
[2]Laboratory of Behavioral Physiology, University of Los Andes, Mérida, Venezuela
phone: + (58) 0274-2402903, fax: + (58) 0274-2402903, email: {ceballos, paredesj, hernande}@ula.ve

ABSTRACT

In this paper, a specific method for massive Capillary Electrophoresis data analysis based on pattern recognition techniques in the wavelet domain is presented. Low-resolution, denoised electropherograms are obtained by applying several pre-processing algorithms including discrete wavelet transform, denoising, detection of region of interest and baseline correction. The resultant signal is mapped into multi-character sequences exploiting the first derivative information and multi-level peak height quantization. Next, local alignment algorithms are applied on the coded sequence for peak pattern recognition. Finally, Gaussian approximation is performed to assure precise peak-height measurements.

1. INTRODUCTION

Current epidemiologic studies, DNA analysis, high temporal resolution neurochemical experiments, drug of abuse screening and the need to lower medicine costs have compelled the development of high throughput techniques including Elisa, microarray and capillary array electrophoresis. In particular, this last technique has been proved to be a cost-effective, rapid and highly efficient separation method that requires minimal sample volume and relatively simple hardware. The use of this technique, however, generates massive amounts of electropherograms demanding data analysis that is mostly done by visual inspection or through human assisted software. On this line of work, emerge the need of developing fast and efficient algorithms based on signal processing tools that allow us to analyze the electropherograms in a fast and reliable mode. In particular, pattern recognition is, perhaps, the most needed signal processing tool in order to cluster and classify massive volume of electropherograms. Pattern recognition, however, in capillary electrophoresis (CE) represents a challenge due to the variability inherently observed in CE data [1].

An electrophoretic register can be thought of as a temporal series composed of a linear superposition of several Gaussian-like waveforms whose temporal location may change due to the migration time shift present in this kind of data [2]. Furthermore, each Gaussian waveform is closely related to a specific substance, hence its peak height varies among electrophoretic registers according to the concentration of the corresponding substance.

Part of the goals in the analysis of electrophoresis data is to identify a substance of interest and measure its concentration, process that is done by visual inspection.

To the best of our knowledge, very little work has been devoted in the development of an automatic system for processing massive electrophoretic data. The closest works related to this research have been reported in [3, 4, 5]. In [3], several post-processing methods are proposed for high-throughput analysis of separation data, it includes polinomial baseline correction, automatic peak marking based on first derivative, aided lineal temporal normalization and assisted deconvolution of peaks. Szymanka et. el in [4] uses dynamic time warping to correct the migration time shifts commonly presented in EC data. Furthermore, a similarity score between two whole-electropherograms is used as a match metric. Finally, in [5], a local alignment algorithm is introduced to deal with non-linear time-shifting and pattern recognition.

Although the approach developed in [3] tries to exploit the preprocessing stages of a possible more complete massive data processing system, in this paper, we propose specific methods to efficiently deal with automatic massive analysis of capillary electrophoresis data including peak pattern recognition. We propose dynamic programming concepts [4, 5] applied in wavelet domain reducing thus computational complexity. Furthermore, the proposed method introduces a robust and versatile baseline correction algorithm, peak pattern recognition for analysis and classification and gaussian approximation for high precision measurements in massive electrophoretic data.

The proposed approach comprises four stages: in a first stage, the electropherograms are pre-processed in the wavelet domain. Data reduction, denoising and region of interest detection are suitably achieved based on the wavelet transform. Secondly, baseline correction algorithm based on second derivative is developed to remove the time-varying offset present in an electrophoretic register. Next, peak pattern recognition by local alignment and automatic gaussian approximation algorithms are applied on the preconditioned registers. All of these stages have been designed to work in automatic mode.

The proposed approach was tested on the analysis of intracerebral microdialysate data, achieving a correct detection rate around 85% with a processing time of less than 0.3 sec-

ond per 25.000-point electropherogram. For a detailed description of the pattern recognition technique see [5].

2. METHODS

Figure 1 shows the four stages of the proposed approach.

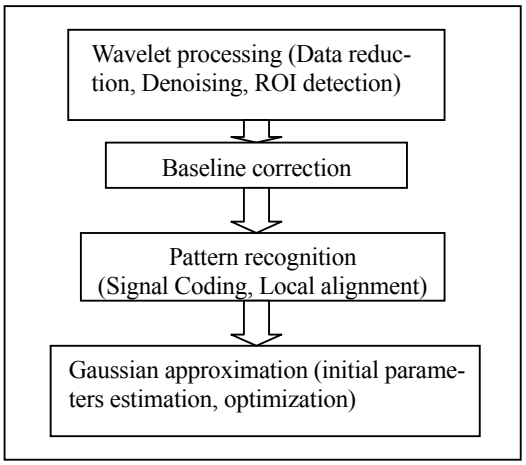

Figure 1 – Stages of the proposed method.

2.1 Wavelet Processing

2.1.1. Selecting Wavelet Type and Resolution Level

Weidong et. al [6] tested several mother wavelets and chose symlet 4 as the one that has the best performance for the analysis of electrophoretic data since it yields the best peak reconstruction and peak preservation. We profited on this research and used symlet 4 in this work.
For pattern recognition we use the approximation coefficients yielded by a four-level wavelet decomposition. At this level of wavelet decomposition, the shape of the original signal as well as the most salient information are preserved. Furthermore, the number of data points is reduced by a factor of 24 and a smoothing operation is applied on the signal by the successive low-pass filtering inherent in the wavelet decomposition.

2.1.2. Detection of Region of Interest (ROI)

We observed that at 7-level wavelet decomposition the detail coefficients have information related to the signal shape and not to the noise components. Therefore, using this information, the region containing peaks of interest is found according to the following procedure: The absolute value of the detail wavelet coefficients at the seven-level wavelet decomposition are calculated, and the region of interest is defined starting at the most left coefficient whose magnitude value is greater than a given threshold value and ending at the right most coefficient with magnitude greater than the threshold value. The corresponding starting and end points of ROI are then found at 4^{th} level wavelet decomposition. We set the threshold value to 2% of the maximum absolute value of the detail wavelet coefficients at level seven.

2.1.3. Noise Reduction

Since pattern recognition is performed on a low resolution signal, a denoising operation is intrinsically applied on the signal. More precisely, the pattern recognition is performed at four-level wavelet decomposition, therefore the wavelet detail coefficients at first four levels are throwed away leading thus to high-frequency noise suppression. Furthermore, the remaining signal noise components are reduced by a thresholding operation performed on detail coefficients of the 5th and 6th wavelet decomposition levels followed by an inverse wavelet transform until the fourth level [5].

2.2 Baseline correction

The baseline of an electropherogram is not very often a perfect horizontal line. Indeed, it can be regarded as an offset that changes dynamically with time being (possibly) unique for each acquired CE register. In some electropherograms, the baseline gradually raises due to spurious fluorescent material adhered to the outer surface of the capillary during the injection procedure. Sometimes, the baseline drifts downwards due to tailing of a highly concentrated band. This baseline drifts makes much harder the pattern recognition problem with methods that use peak amplitudes to compare patterns. Furthermore, it may lead to wrong peak measurements due to the offset introduced on the data points. Therefore, baseline correction emerges as a required pre-processing stage before any further downstream data analysis.
In the present study, we exploit the fact that baseline is a low frequency signal and, therefore we analysis the four level wavelet approximation coefficients to estimate the baseline curve heuristically. The algorithm is based on cubic interpolation using key preselected wavelet coefficients as described next.
Let Y=[$Y(1), Y(2), \ldots, Y(M)$] and B=[$B(1), B(2), \ldots, B(M)$] be a first approximation of the baseline curve and the final baseline curve to be estimated, respectively where M is the number of wavelet approximation coefficient at four decomposition level. Define each entry of Y as follows:

$$Y(k)=\begin{cases} C(k) & if \quad |Z(k-i)|<T(k-i) \quad for \quad i=1,0,-1 \\ 0 & Otherwise \end{cases} \qquad (1)$$

Where $Z(n)$ is the second derivative of the wavelet coefficients C(k), and $T(k)$ is a threshold function equal to a rescaled version (between h_{min} and h_{max}) of a function constructed by linear interpolation using de local maximum of $C(k)$, where h_{min} and hmax are tunning parameters $0< h_{min} < h_{max} < 20$.
Replace those components of Y, defined by the "otherwise" part of the equation (1), by cubic interpolation using those elements of Y for which $Y(k)=C(k)$.
Define each component of B as follows

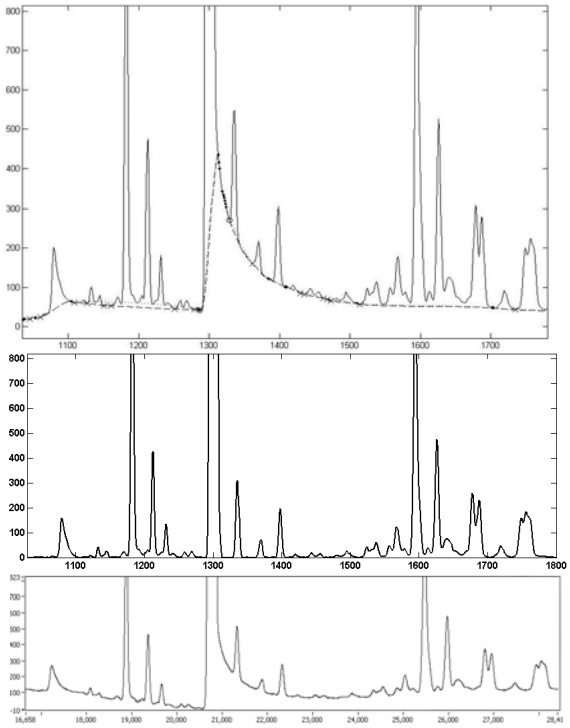

Figure 2 – Construction of the baseline curve. Upper: Dashed line: Final baseline curve, Dotted line: First approximation of the baseline curve. • wavelet coefficients with smooth variations, × local minimum under the first approximation of the baseline curve. Middle: baseline correction by the proposed approach. Bottom: baseline correction achieved by [3].

$$B(k)=\begin{cases} C(k) & \text{if} \quad \begin{array}{c} |Z(k-i)|<T(k-i) \quad \text{for} \quad i=1,0,-1 \quad \text{or} \\ C(k)<Y(k) \quad \text{and} \quad C(k) \quad \text{is a local minimum} \end{array} \\ 0 & \text{Otherwise} \end{cases} \qquad (2)$$

Replace those components of *B*, defined by the "otherwise" part of equation (2), by cubic interpolation using those elements of *B* for which *B(k)=C(k).*

As can be seen from Equations (1) and (2), the baseline curve is forced to pass through those wavelet coefficients that have smooth variation. Figure 2 shows the baseline constructed using the proposed algorithm. Note that the local minimum, marked as x, that are below the first approximation of the baseline curve are taken into account to define the final baseline curve. We found this method for baseline correction more suitable in data with fast variation in baseline than polynomial methods. Figure 2 (middle) shows the performance of our methods for baseline correction in a particular case with fast drift on the baseline. Figure 2 (bottom) depicts the baseline correction yielded by the approach described in [3] where singular valued decomposition and a 10 degree polynomial function is used to model the baseline.

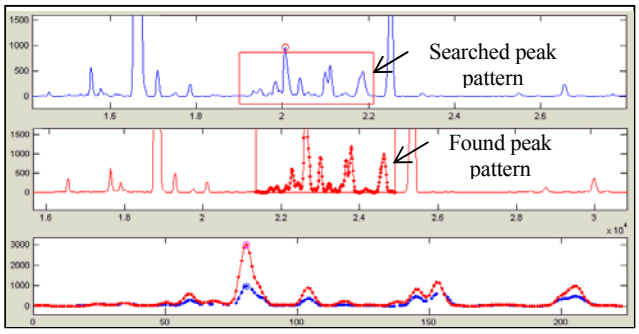

Figure 3 – Local alignment of electropherograms performed on the approximate wavelet coefficients at fourth wavelet decomposition level.

2.3 Pattern Recognition

The resultant denoised electropherogram at the four-level wavelet decomposition with the baseline removed is coded using a finite alphabetical codes where the codewords are codes associated with not only to the first derivative signs like in [7] but also to the height of the peaks.

2.3.1. The proposed coding method

The proposed coding method is as follows: each point in the electropherogram is coded as M if the signal slope is positive, P if the signal slope is negative, L if the point is a valley and either A, B, C, D, E, F, G, H if the signal reaches a local maximum. In this last case, the assigned alphabetical code depends on the height of the peak. Thus, the A to H characters correspond to peak height quantized to eight levels uniformly distributed between 0 and 4000 milivolt.

2.3.2. Local Alignment

Having coded the desired pattern and the electropherograms at low resolution, we applied the pairwise local alignment method of Smith & Waterman [8] to find the desired coded-pattern in each coded electropherogram. To achieve that, we use a substitution matrix, a gap opening penalty and a gap extension penalty described in full details in [5].

Figure 3 depicts a local alignment achieved with these parameters. Note that the algorithm successfully finds the desired pattern even though there exist two aligned peaks with a variation in peak height of more than 100%. Note also that, in finding the desired pattern, the algorithm aligns both the searched pattern and the found pattern, by inserting several gaps in suitable locations.

The gap insertion and the ability of aligning peaks with remarkable height difference are the main advantages of the proposed approach to handle the variability found in electrophoretic signals. This mode of application of dynamic programming (local alignment) for peak pattern matching is different from [4] where the whole electropherograms are compared.

2.3.3. Optimization of the Local Alignment Algorithm

An approach to speed up the Smith & Waterman alignment algorithm consists in searching for the highest alignment score only in the last row of the DP matrix instead of searching in the whole matrix. Due to the high variability found in the electropherograms, the only solution outputted by the conventional Smith & Waterman algorithm may lead to a wrong detection (misplace) of the desired pattern. To overcome this drawback, we consider several possible solutions associated with different time locations of the found pattern, giving preference to the pattern with the closest location to the location of the desired pattern in the reference electropherogram.

Let P_0 denotes the location of the desired pattern in the reference electropherogram, and $\Delta p/2$ a tuneable parameter that represents the maximum deviation around P_0 for the allowed solutions. The parameter ($\Delta p/2$) can be chosen based upon the peak location reproducibility of the electropherograms. In this work, we set $\Delta p = 0.14P_0$ based upon the maximal deviation observed in the tested data. Furthermore, as possible solutions, we consider the five highest alignment score that are related to five different locations in the electropherogram.

We consider the alignment scores of the last raw (on the dynamic programming matrix) in descending order and apply the backtracing algorithm [5] repeatedly until 5 different locations of the found pattern are obtained. In this process, if one of the optimal reconstructed pathways ends in a location inside the preference band (P_0-$\Delta P/2$, P_0+$\Delta P/2$), this reconstruction will be considered as the final alignment solution and not additional backtracing is performed. On the other hand, if none of the 5 solutions that have been considered lies inside the band, the alignment that starts closest to P_0 will be considered as the solution.

2.4 Gaussian approximation

The superposition of adjacent substances inner the capillary can lead to erroneous measurements of peak heights. It is necessary the deconvolution of signal in kernel functions. Note in Fig. 4 decomposition of electropherogram curve in sum of gaussian kernels.

If the distance between the detection cell and detector is short and wall desorption is negligible, it is expected that the waveform related to a particular substance has a Gaussian-like shape mainly due to longitudinal diffusion of substances inside the buffer [2]. We approximate the peaks by sum of Gaussian functions using nonlinear optimization algorithm for estimation of the parameters (peak height, location and variance) of a Gaussian waveform (Trust Region Algorithm). The initial parameters are automatically calculated exploiting the information related to the second derivative of electropherograms. More precisely, the zero-crossing of the second-derivative represents a rough estimation for the variance, whereas the average of two successive zero-crossing of the second-derivative defines the initial gaussians location and the amplitudes are calculated by linear regression [9]. Perhaps the most of works in deconvolution of peaks in capillary zone electrophoresis uses second derivative [10], if 4th derivative is considered instead 2nd , and the peaks are assumed gaussians, we can resolve high level overlapping of substances. Note the high level overlapping resolved in the sum of the left most second and third Gaussians in left image in Fig.4. Unlike [9], where the Gaussians' amplitudes are allowed to take on negative values, in our approach, only positive Gaussians are considered. The Gaussian processing time was 90 sec. by 120 gaussians approximately. The process is performed in third level of wavelet decomposition in order to conserve a suitable resolution of peaks. Note in Fig.4 the approximation on level 3 and 4 of wavelet decomposition.

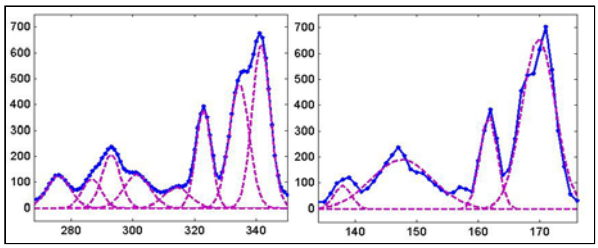

Figure 4 – Approximation of an electropherogram by linear combination of Gaussian functions at approximation wavelet coefficients. Left: third decomposition level, right: fourth decomposition level.

Once all the electropherograms have been processed and the found patterns have been aligned by the peak matching procedure described above, the results can be reorganized in a matriz that can be analized by datamining algoritm with the aim of finding no explicit information.

3. RESULTS AND DISCUSION

3.1 Analysis of Intracerebral Microdialysate

We randomly select 30 electropherograms out of a set of 277 electropherograms obtained from dialysates of the same rat's brain area. Patterns containing 4, 8, 12, 16, 32 and 50 peaks are selected in an arbitrary reference electropherogram chosen out of the set. Those patterns are then searched in the subset of electropherograms. Table 1 shows the percent of correct detection as a function of the length of the pattern. The percent of correct detection is validated by expert visual inspection, a correct detection is assumed when all peaks in searched pattern are aligned with corresponding peaks in found pattern.

Tabla 1. Percent of correct detection achieved by the proposed algorithm.

Pattern length (number of peaks)					
4	8	12	16	32	50
65.3%	84%	85.3%	80.6%	79.5%	90%

As can be noted in Table 1, the proposed method yields competitive results for longer pattern.

As a second test, the proposed algorithms are applied in the whole data set (277 electropherograms) on a Pentium IV, 3.2

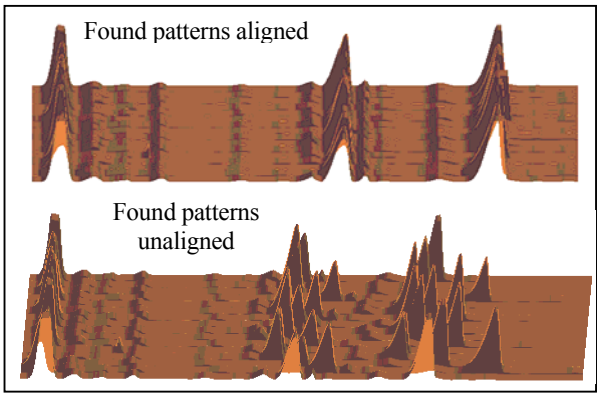

Figure 5 – Three-dimensional surfaces representing found patterns. Top: unaligned patterns. Bottom: aligned patterns.

GHz, with 1GB RAM. For an eight-peak pattern the coding and local aligning time take just 23 ms per electropherogram, whereas the signal conditioning time is about 64 ms per electropherogram.

To further illustrate the performance of the proposed algorithm, Fig. 5 shows the results of the alignment process using a 3-dimensional representation. In that representation, each row is a found pattern, hence, a set of consecutive columns are associate with chemical substances. The lower image shows the found pattern without being aligned, whereas the upper image shows the found patterns aligned with the reference pattern. Interestingly, this kind of representation of the aligned data leads to a rapid visual evaluation of the found patterns in a sequence of electropherograms. Moreover, the surface allows us to detect those peaks showing the largest variations and, therefore, the chemical changes among the electropherograms.

Finally, Fig. 6 shows another set of found patterns that, by convenience, have been sorted according to the instant of sample acquisition. In this representation de third dimension which represents relative substance concentration is mapped to colours. Note a notable increment on the concentration of a certain chemical substance (second substance, from left to right) in the time course (vertically direction). This variation may be produced by a specific experimental manipulation such as the study of the effect of a drug in variation of concentrations of a certain set of substances.

4. CONCLUSIONS

In this paper, a pattern recognition approach for electrophoretic data processing has been proposed that achieves about 85% of correct pattern detection and an execution time less than 0.3 seconds per 25000-point electropherogram. This percentage of correct detections tends to improve as the length of the searched pattern increases.

The proposed methodology can certainly have a great impact on modern high throughput capillary electrophoresis instrumentation. In this particular area, the methods presented in this and future articles will substitute slow, human based time-consuming visual pattern recognition methods by automatic fast pattern recognition techniques.

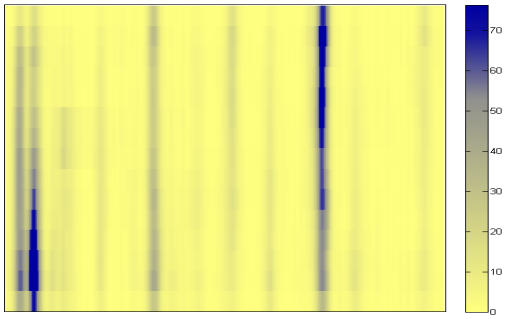

Figure 6 – Variations in the concentration of chemical substances in the tested sample as the experimental conditions changes. All the aligned electropherograms are assembled into a matrix so that each raw is an aligned electrophoretic register. Right: colorbar indicating substance relative concentration level. Low concentration (~0), high concentration (~ 70).

REFERENCES

[1] Schaeper J., Sepaniak M., "Parameters affecting reproducibility in capillary electrophoresis". *Electrophoresis*, vol. 21(7), pp.1421-1429, 2000.

[2] Gebauer P. and Bocek P. "Predicting Peak Symmetry in Capillary Zone Electrophoresis: The Concept of the Peak Shape Diagram". *Analytical Chemistry*, vol. 69(8), pp.1557-1563, 1997.

[3] Shackman J., Watson C., Kennedy R., "High-throughput automated post-processing of separation data", *Journal of Chromatography A.*, vol. 1040, pp. 273-282, 2004.

[4] Szymanka E.,Markuszewski M., and others., "Evaluation of different warping methods for the analysis of CE profiles of urinary nucleosides", *Electrophoresis*, vol. 28, pp. 2861-2873, 2007.

[5] Ceballos G., Paredes J., HErnández L., "Pattern recognition in capillary electrophoresis data using dynamic programming in the wavelet domain.", *Electrophoresis,* vol. 29, pp. 2828-2840, 2008.

[6] Weidong C., Xiaoyan C., Xiurong Y., Erkang W., "Discrete wavelets transform for signal denoising in capillary electrophoresis with electrochemiluminescence detection", *Electrophoresis,* vol. 24(18), pp. 3124–3130, 2003.

[7] Guillo C., Barlow D., Perrett D., Hanna-Brown M., "Micellar electrokinetic capillary chromatography and data alignment analysis: a new tool in urine profiling", *J Chromatogr A* , vol. 1027(1-2), pp. 203–212, 2004

[8] Smith T., Waterman M. ,"Identification of common molecular subsequences". *J Mol Biol*, vol. 147(1), pp. 195–197, 1981.

[9] Goshtasby and O'Neill W., "Curve Fitting by Sum of Gaussians". *CVGIO:Graphical Models and Image Processing*, vol. 56(4), pp. 281-288, 1994.

[10] Olazábal V., Prasad L and others, "Application of wavelet transform and an approximate deconvolution method for the resolution of noisy overlapped peaks in DNA capillary electrophoresis", *Analyst*, vol. 129, pp. 73-81, 2004.