CORRECCIÓN DE LÍNEA BASE EN DATOS ELECTROFORÉTICOS USANDO OPTIMIZACIÓN LOCAL DEL ALGORITMO LEGEND EN EL DOMINIO WAVELET

JOSÉ L. PAREDES y ENEDINA SOSA

RESUMEN

Se propone un método de corrección de la línea base (LB) de señales electroforéticas que explota la representación wavelet a baja resolución de la señal original. La LB es modelada en el dominio wavelet como una función polinomial y se usa el algoritmo de optimización LEGEND para determinar los parámetros del modelo polinomial que mejor se ajusta a una subsección de la LB, de forma tal de minimizar una función costo asimétrica robusta. El algoritmo propuesto produce una corrección adecuada de la LB en aquellas zonas del electroferograma donde se aprecian sustancias de baja concentración en las adyacencias de sustancias de concentración elevada, preservando picos asociados con las primeras. Se validó el algoritmo desarrollado en un problema de medición de la cantidad de glutamato presente en 24 registros electroforéticos y se comparó su desempeño con los valores medidos por el especialista donde la línea base es corregida en forma manual y con los valores arrojados por un segundo algoritmo de corrección de LB recientemente propuesto.

La electroforesis capilar es una técnica de alta resolución para separación de sustancias químicas que permite obtener información precisa acerca de los componentes que conforman dos matrices químicas complejas. Su potencial de aplicación cubre distintas áreas, que incluyen bioanálisis, biotecnología, las industrias farmacéutica (Gamero, 2001), química y de alimentos (Alexandra *et al*., 2003), entre otras. Como en otras técnicas de separación de compuestos químicos, la interpretación de los electroferogramas implica un proceso de reconocimiento de patrones que representan las sustancias de interés. Este proceso es realizado generalmente en forma visual dada la poca reproducibilidad observada en las señales, que dificulta la aplicación de técnicas automáticas de reconocimiento de patrones.

Un registro electroforético puede ser considerado como una serie temporal conformada por la superposición de tres tipos de señales. Primero, la señal que representa los componentes químicos que conforman la muestra en estudio y que son de interés para el especialista; dicha señal puede ser modelada como la superposición de formas de onda similares a una campana de Gauss cuya amplitud máxima (pico) representa la masa de la sustancia química presente en la muestra analizada, su localización y el tiempo de migración (Shanle *et al*., 1997). La segunda señal presente es la que refleja las limitaciones técnicas del proceso de adquisición, que se manifiestan como componentes de ruido aleatorio, generalmente caracterizado por una distribución gaussiana. Finalmente, una señal indeseada que representa la influencia de diversos procesos físicos concurrentes con la electroforesis capilar; esta última componente se manifiesta como una señal de baja frecuencia denominada línea base (LB), que tiende a introducir error en la medición de los picos, en especial cuando hay picos de baja altura en las adyacencias de picos elevados.

A fin de ilustrar los inconvenientes causados por la presencia de la LB, en la Figura 1 se muestran los distintos componentes de la señal electroforética. En el caso del pico ubicado en las proximidades del instante de migración 90, a la derecha del pico más elevado, al no considerarse la corrección de LB, la medida es de ~570, y al hacer la corrección manual (substracción manual de la señal línea base a la señal electroforética original) el valor del pico seria ~370, lo cual corresponde a un error del 54,05%, que es considerablemente elevado y pudiera llevar al especialista a una mala interpretación de los análisis realizados. Surge así la necesidad de desarrollar algoritmos de corrección de LB que permitan suprimir o eliminar esta señal indeseada del registro electroforético.

La corrección de línea base previo el reconocimiento de patrones electroforéticos ha sido recientemente tratada en Ceballos *et al.* (2007, 2008). Estos autores abordaron el proceso de corrección de LB usando

PALABRAS CLAVE / Algoritmo LEGEND / Corrección de Línea Base / Electroforesis Capilar / Transformada *Wavelet /*

Recibido: 22/07/2008. Aceptado: 10/08/2009.

José L. Paredes. Ph.D. en Ingeniería Eléctrica, University of Delaware, EEUU. Profesor, Universidad de Los Andes (ULA), Venezuela. Dirección: Grupo de Ingeniería Biomédica, Escuela de Ingeniería Eléctrica, ULA, Mérida 5101, Venezuela. e-mail: paredesj@ula.ve

Enedina Sosa. Ingeniera Electricista, ULA, Venezuela. Ingeniera Inspectora, Instituto Merideño de Desarrollo Rural, Venezuela. e-mail: enedina@ula.ve

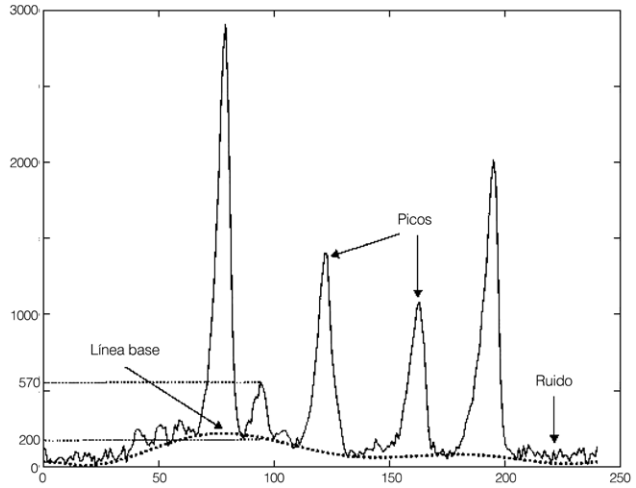

Figura 1. Componentes de una señal electroforética. Influencia de la línea base en la medición de la concentración de una sustancia.

métodos no paramétricos, donde la estimación de la señal LB se hace mediante el estudio de concavidad, así como la rapidez de cambio de los coeficientes de aproximación *wavelet* a una determinada resolución, consiguiéndose ciertos puntos críticos de la LB. Seguidamente, un proceso de interpolación cúbica permite estimar todos los puntos de la LB. Sin embargo, este método suprime picos pequeños asociados con sustancias que contienen información relevante acerca de los componentes químicos que conforman la muestra.

En el presente trabajo se propone un esquema de corrección de LB en el dominio *wavelet*, donde la estimación de la señal LB y su posterior corrección se realiza usando una señal electroforética de baja resolución y de reducidos componentes de ruido. La Figura 2 muestra el diagrama de bloque de la corrección de LB propuesta. Como se puede apreciar, la señal original es previamente acondicionada usando técnicas de procesamiento multi-resolucional (pre-procesamiento *wavelet*) con el fin de reducir las componentes ruidosas y disminuir la resolución del registro electroforético original. La salida del bloque conforma la señal fundamental que se utiliza en el proceso de estimación y corrección de LB.

Seguidamente, se modela la línea base como una función polinomial y se utiliza el algoritmo de optimización LEGEND para determinar los parámetros de dicho modelo, de forma tal de minimizar una función costo asimétrica robusta. El algoritmo propuesto extiende el uso del algoritmo LEGEND, propuesto recientemente (Mazet *et al*., 2005, 2004) para corregir la LB en espectros infrarrojos, a la estimación de la LB en datos electroforéticos. A diferencia del algoritmo original, que resuelve un problema de optimización global, el método propuesto corrige localmente la LB en el dominio *wavelet*, resultando en un mejor desempeño. Se observó que el algoritmo propuesto produce una corrección adecuada de la LB, en particular en aquellas zonas del electroferograma donde aparecen picos de baja altura en las adyacencias de picos elevados. El desempeño del algoritmo desarrollado fue comparado con el desempeño del algoritmo propuesto por Ceballos *et al*. (2007, 2008) en la medición de las masas o cantidad de glutamato en un conjunto de 24 electroferogramas. Se mostrará que el algoritmo de corrección de la LB propuesto produce menores errores en la medición de la cantidad de glutamato, además de preservar la mayoría de los picos, aún los pequeños.

En adelante se da una breve descripción de la transformada *wavelet* y sus aplicaciones, que involucra reducción de resolución y disminución de ruido, y luego se describe el modelo matemático usado para representar el registro electroforético y la LB en el dominio *wavelet*. Tras plantear el problema de corrección de la LB desde un punto de vista de minimización de una función costo asimétrica, se justifica el uso de la función costo seleccionada y se describe el algoritmo iterativo empleado para su minimización. Así mismo, se presenta una comparación entre el método propuesto y el descrito por Ceballos *et al*. (2007, 2008). Finalmente, se presentan las conclusiones alcanzadas en el desarrollo de este trabajo.

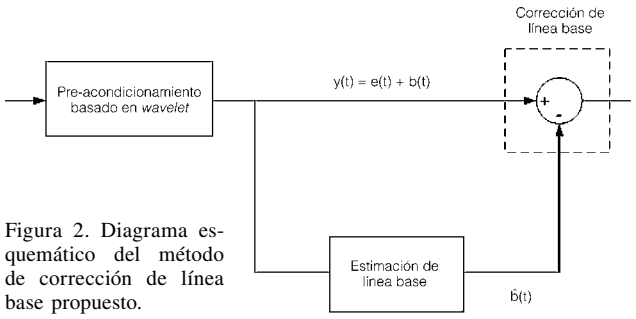

Figura 2. Diagrama esquemático del método de corrección de línea base propuesto.

Pre-procesamiento *wavelet*

La transformada *wavelet* es una poderosa herramienta para el análisis de señales unidimensionales y bidimensionales pues permite la caracterización simultánea, tanto en tiempo como en frecuencia, de transientes o comportamientos no estacionarios en la señal de interés. Su potencial de uso se ha extendido recientemente para el análisis de señales provenientes de electroforesis capilar, específicamente para la limpieza de registros electroforéticos (Perrin *et al*., 2001; Weidong *et al*., 2003; Weiping *et al*., 2006) y el reconocimiento de patrones en electroforesis capilar (EC; Ceballos *et al*, 2007, 2008). A continuación se describe brevemente el principio de la transformada *wavelet* discreta y el uso de la misma para la supresión de componentes ruidosas en una señal.

Transformada wavelet discreta

Sea X(t) una señal en tiempo discreto de tamaño 2^n que pudiera representar la señal adquirida por un equipo de electroforesis, muestreada a una frecuencia f_s. Aplicar la transformada *wavelet* discreta sobre la señal en estudio implica hacerla pasar por un banco de filtros digitales en cascada con operaciones de submuestreo, obteniéndose a la salida dos componentes. Una primera componente, conocida como coeficientes de aproximación *wavelet*, contiene las componentes de baja frecuencia de la señal de entrada, $(0,f_s/4$, y no es más que una representación de la señal en estudio a baja resolución. La segunda componente, conocida como coeficientes de detalles, contiene las componentes de alta frecuencia de la señal en estudio $(f_s/4, f_s/2)$. En estas componentes se encuentran los detalles que caracterizan a la señal de entrada así como también las componentes de ruido que se suman inevitablemente durante el proceso de adquisición. Analíticamente, la transformada *wavelet* discreta de la señal X(t) se reduce a

$$c(t)= \sum_k g(k)X(k+2t) \quad (1)$$

$$d(t)= \sum_k h(k)X(k+2t) \quad (2)$$

donde g(k) y h(k): coeficientes del filtro paso bajo y paso alto, relacionados con las funciones escalar y *wavelet*, respectivamente (Sidney *et al*., 1998). Como se puede observar en las Ecs. 1 y 2, además de la operación de convolución típica en un proceso de filtrado, existe una operación de submuestreo implícito. Es decir, solo una de cada dos muestras se mantiene a la salida del filtro. Esta operación de submuestreo no produce pérdida de información relevante de la señal, dado que las señales a la salida del filtro contienen la mitad de la frecuencia de la señal de entrada, por consiguiente su frecuencia de muestreo puede reducirse a la mitad. Así se evita redundancia en la representación, consiguiéndose a la salida de la descomposición *wavelet* los coeficientes de aproximación c(t) y de detalles d(t), de tamaño igual a la mitad del tamaño de la secuencia de entrada, es decir 2^{n-1}.

Las Ecs. 1 y 2 definen un primer nivel de descomposición de la transformada *wavelet*. Niveles sucesivos de des-

composición se obtienen al aplicar progresivamente las operaciones de filtrado sobre la señal que resulta a la salida del filtro paso bajo. Específicamente.

$$c_{j+1}(t)= \sum_k g(k)c_j(k+2t)$$

$$d_{j+1}(t)= \sum_k h(k)c_j(K+2t)$$

donde c_j y d_j: coeficientes *wavelets* de aproximación y de detalles en la j-enésima escala, respectivamente. En la medida que se avanza en la descomposición se consiguen representaciones de la señal de entrada a distintas resoluciones. Así, para j= 0 se tiene la resolución original de los datos de entrada, es decir $c_0(t)$= X(t). A medida que j aumenta la resolución disminuye progresivamente. A esta descomposición sucesiva se le conoce como análisis multi-resolucional (Sidney *et al*., 1998). La Figura 3 muestra el esquema de la transformada *wavelet* discreta mediante un proceso de filtrado sucesivo, presentándose, en este caso, tres niveles de descomposición. El símbolo 2↓ denota la operación de submuestreo por un factor de dos, CA_j y CD_j representan los coeficientes de aproximación y de detalles, respectivamente, al j-enésimo nivel de descomposición.

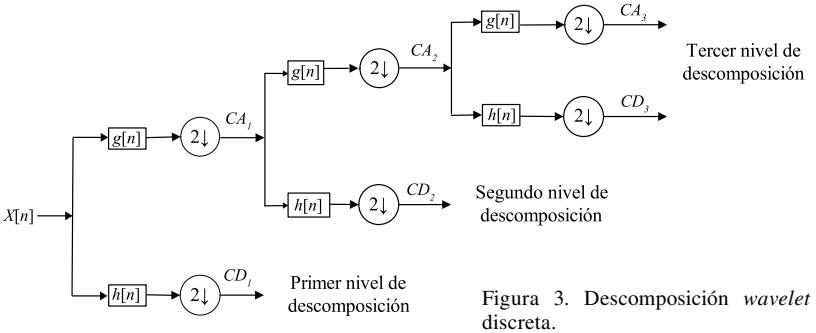

Figura 3. Descomposición *wavelet* discreta.

Al igual que otras transformadas, la transformada *wavelet* es reversible; es decir, a partir de los coeficientes *wavelets* de aproximación c(t) y de detalles d(t), es posible reconstruir la señal de entrada original mediante operaciones de filtrado paso bajo y paso alto. Esto es cierto si la *wavelet* satisface la propiedad de reconstrucción perfecta, condición que impone ciertas restricciones a los coeficientes de los filtros paso alto y paso bajo (Sidney *et al*., 1998). En forma más general, la señal en la j-enésima resolución puede reconstruirse a partir de los coeficientes de detalles y de aproximación de la resolución (j + 1)-enésima de la siguiente forma

$$c_j(n)=\sum_k c_{j+1}(k)g_l(n-2k)+d_{j+1}(k)h_l(n-2k) \quad (3)$$

donde $g_l(n)$ y $h_l(n)$: coeficientes de los filtros de reconstrucción paso bajo y paso alto, respectivamente.

Limpieza de señales usando la transformada wavelet discreta

Una de las aplicaciones de mayor interés de la transformada *wavelet* es la supresión de las componentes ruidosas que contaminan a la señal en estudio. El principio de esta metodología de limpieza de señales es el hecho que las componentes de alta frecuencia de la señal (detalles y ruido) se encuentran concentradas en los coeficientes de detalles de la descomposición *wavelet* (altas frecuencias). Por ello, si estos coeficientes se modifican, por ejemplo se hacen cero si su valor absoluto es menor que un cierto valor umbral o, de lo contrario, se dejan inalterados, en el proceso de reconstrucción la señal obtenida es una versión limpia de la señal contaminada. La escogencia del valor umbral apropiado para cada aplicación es motivo de interés (Donoho, 1995; Perrin *et al*., 2001; Weidong *et al*., 2003; Weiping *et al*., 2006), pues un valor de umbral muy elevado elimina detalles de interés que para la actual aplicación pudieran ser picos de pequeña amplitud asociados con bajas concentraciones de sustancias, mientras que un valor muy bajo del umbral deja componentes ruidosas en la señal reconstruida, dificultándose el proceso de estimación en la LB. En resumen, la operación de limpieza se reduce a aplicar la operación de reconstrucción dada por la Ec. 3, sustituyendo los coeficientes de detalles d_j por una versión modificada de los mismos, la cual pudiera ser la formulación propuesta por Donoho (1995),

$$\hat{d}_j= U\left(d_j,u\right)= \begin{cases} d_j & si \ \left|d_j\right|> u \\ 0 & si \ \left|d_j\right| \le u \end{cases} \quad (4)$$

siendo u el umbral a ser óptimamente escogido.

En el presente estudio la transformada *wavelet* es usada para reducir la resolución del electroferograma original, así como para disminuir las componentes de ruido.

Modelo de la Señal Electroforética

La señal pre-acondicionada luego de la etapa de procesamiento *wavelet* se puede modelar como

$$y(t)= e(t) + b(t)$$

donde e(t): señal residuo, incluyéndose aquí los picos que representan la concentración de los componentes-químicos de la sustancia en estudio, y el remanente del ruido y las incertidumbres presentes en los datos para el instante de migración t; y b(t): señal LB que se desea estimar para su posterior remoción del registro electroforético.

El origen de esta señal b(t), en el caso de la detección mediante fluorescencia inducida por láser, se debe a múltiples fenómenos físicos/químicos, entre los que destacan una elevada concentración de algunas sustancias presentes en la muestra analizada y ruido proveniente de diversas fuentes, tales como luces de fuentes externas (lámparas y fluorescentes), calidad y cantidad del material fluorescente usado en el proceso electroforético, fluorescencia de las paredes del capilar, variaciones en la intensidad del láser, fotodetectores no bien conectados a tierra, corriente oscura del fotodetector, radiación errante originada por múltiples rebotes del láser dentro de la caja óptica, y el *offset* debido al envejecimiento de los componentes electrónicos de la tarjeta de adquisición.

En todo proceso de corrección de línea base, primero, se requiere estimar la función LB, b(t), que se encuentra oculta en los datos adquiridos, para posteriormente sustraerla de la señal original, tal como se muestra en la Figura 2. Sin embargo, entre los retos que se presentan al momento de desarrollar algoritmos de corrección de LB se encuentran la preservación de los picos pequeños que contienen información relevante para los especialistas del área, así como también la eliminación de las contribuciones no deseadas en la medida de los picos debido a la presencia de la LB. Por consiguiente, una sobre-estimación de la LB conduce a la eliminación de los picos pequeños (sustancias con baja concentración), mientras que una sub-estimación deja residuos que inducen errores en la medición de la concentración de los componentes químicos de la muestra en estudio tal como se señaló arriba.

En este trabajo se considera la señal LB como una señal de baja frecuencia modelada por una función polinomial de orden p. El hecho de usar una función polinomial para modelar la LB se debe a que de acuerdo a Mazet *et al*. (2005) dicha función caracteriza apropiadamente diversas líneas bases presentes en distintos espectros. Sin embargo, el definir un modelo matemático más elaborado de la LB en un electroferograma es un tema abierto de estudio. En el presente trabajo se usa el modelo polinomial pues permite simplificar el proceso de optimización de los parámetros que lo conforman y a su vez presentan un buen desempeño. Así, la LB se modela como

$$b(t)= a_0 + a_1t + a_2t^2 + a_3t^3 + \ldots + a_pt^p$$

donde los a_i para i= 0,1,…,p: parámetros del modelo a ser determinados en forma óptima.

Por conveniencia, se denotará el registro electroforético y la LB como vectores columnas. Esto es, $\mathbf{y}= (y_1, y_2, \ldots, y_N)^T$, siendo la i-enésima componente del vector **y** el valor del registro electroforético en el instante de adquisición t_i, N: número de puntos del registro electroforético, y **T**: operador de transposición. Similarmente, $\mathbf{b}= (b_1, b_2, \ldots, b_N)^T$ donde $b_i= b_{(t)}$.

Usando esta notación, la LB se puede expresar como **b= Ta**, donde **T**: matriz de Vandermonde del vector tiempo, definida como

$$\mathbf{T}=\begin{bmatrix} t_1^0 & t_1^1 & \cdots & t_1^p \\ t_2^0 & t_2^1 & \cdots & t_2^p \\ \vdots & \vdots & & \vdots \\ t_N^0 & t_N^1 & \cdots & t_N^p \end{bmatrix}$$

y $\mathbf{a}= (a_0, a_1, \ldots, a_p)^T$: vector de los coeficientes del polinomio cuyos valores son óptimamente obtenidos siguiendo el algoritmo que se describe en la siguiente sección.

Estimación de la Función LB

El objetivo principal que se persigue es determinar los coeficientes del modelo polinomial de tal forma que éste se ajuste a la señal LB para posteriormente sustraerla. Esto, inevitablemente, conlleva a un proceso de optimización, donde los parámetros del modelo se obtienen minimizando una función costo de la forma (Mazet *et al*., 2005)

$$J(\mathbf{a})= \sum_{i=1}^{N} \varphi(y_i-b_i)= \sum_{i=1}^{N} \varphi(y_i-(\mathbf{Ta})_i) \qquad (5)$$

donde $(\mathbf{Ta})_i$: i-enésima componente del vector **b= Ta**, y_i: *i*-enésima muestra del registro electroforético, y $\varphi(\cdot)$: función costo a ser definida más adelante. Así, los coeficientes a_0, a_1, a_2,…a_p se obtienen de forma tal que cierta medida de distancia, dada por $\varphi(\cdot)$ entre los datos reales y_i, y el modelo propuesto es minimizada.

Diseño de la función costo φ(x)

Un primer método de estimación de los parámetros del modelo de la LB surge naturalmente si se utiliza el enfoque clásico de mínimos cuadrados. Esto consiste en encontrar los coeficientes **a**, que reducen al mínimo el error medio cuadrático entre la señal y la LB. Así, la Ec. 5 se reduce al método de mínimos cuadrados si $\varphi(x)= x^2$, obteniéndose como coeficientes del modelo

$$\hat{\mathbf{a}}= (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{y} \qquad (6)$$

Sin embargo, este enfoque arroja un resultado poco útil, dado que la función costo, $\varphi(x)= x^2$, da el mismo costo cuadrático a cada valor $y_i-(\mathbf{Ta})_i$. Por consiguiente, valores elevados de diferencia entre los picos del electroferograma y la función LB tienen un costo alto y en consecuencia tratará de desplazar la LB hacia los picos originando que la LB estimada pase por encima de picos pequeños, produciendo señales de valor negativo en el momento de hacer la corrección, lo cual no concuerda con el comportamiento asociado a un electroferograma.

Para tratar este problema se pueden utilizar funciones costo que den un costo más bajo a valores elevados y mantengan un comportamiento cuadrático en la vecindad de cero, es decir, cuando la señal y la LB estén cercanas entre sí. Esto puede lograrse si la función costo crece más lentamente que una función cuadrática o exhibe un efecto de saturación, si la diferencia entre los datos reales y la LB estimada supera cierto umbral k de modo que los picos tengan menor influencia en la estimación de la LB.

En el caso particular de las señales electroforéticas, donde solo se tienen picos positivos, se propone utilizar una función costo asimétrica, tal como la función Huber asimétrica (Huber, 2004; Mazet *et al*., 2005) la cual exhibe un comportamiento lineal cuando la variable independiente toma valores por encima del umbral k. Así, un pico elevado afectará la estimación de la señal LB como si fuera un pico de valor relativamente pequeño. Dicha función costo viene dada por:

$$\forall x \in R, \quad \varphi(x)=\begin{cases} x^2 & \text{si } x<k \\ 2kx-k^2 & \text{en otro caso} \end{cases} \qquad (7)$$

La función costo asimétrica usada, presenta la particularidad de dar un costo bajo a aquellos puntos que se encuentran bastante alejados de la señal estimada, siempre y cuando arrojen una diferencia positiva entre los datos reales y la LB, y un costo cuadrático cuando la diferencia es menor al parámetro k. Para ser más específicos:

- Si la diferencia entre la señal original y la señal LB estimada para un instante t_i toma valores positivos elevados, tal diferencia no influirá severamente en la determinación de los parámetros del modelo **a**, pues dicha diferencia cae en la parte de crecimiento lineal de la función costo φ(x). Así, un pico elevado influirá en la determinación de **a** de igual forma que uno pequeño, hecho que resulta beneficioso para la estimación de LB de las señales electroforéticas.

- Cuando se tiene una diferencia relativamente grande pero negativa, es decir, la LB estimada se encuentra por encima del registro electroforético en un instante de migración t_i, dicha diferencia debería influir fuertemente en la estimación de **a** tratando de desplazar hacia abajo la LB, evitando así que esa parte de la señal se haga negativa en el momento de la corrección de la LB.

- Finalmente, si la diferencia es relativamente pequeña y positiva (o negativa), se encontraría en la parte cuadrática alrededor de cero de la función costo y por tanto su contribución en la estimación de la LB es tratar de desplazarla hacia arriba (o hacia abajo) tratando de pasar por los mínimos locales que presenta el registro electroforético, evitando dejar residuo de la LB (o perder información de la señal de interés suprimiendo picos pequeños) en el momento de corregirla.

Cabe mencionar que otras funciones costo asimétricas tales como la función cuadrática truncada, ampliamente utilizada en estadística robusta, pueden ser usadas. En este trabajo se encontró que la función Huber presentó el mejor desempeño, además de ser diferenciable en todo su dominio.

Optimización por minimización semi-cuadrática local

El objetivo entonces se reduce a resolver el problema de minimización dado por la Ec. 5 usando la función costo definida por la Ec. 7. A diferencia del método de mínimos cuadrados, donde se tiene una expresión cerrada para la estimación de los parámetros del modelo, resolver el problema de minimización de la Ec. 5 usando la función de Huber asimétrica no es directo. Para minimizar este tipo de funciones se usará la minimización semi-cuadrática (*half-quadratic minimisation*; HQ), la cual es una técnica iterativa que simplifica la optimización de un criterio no cuadrático (Mazet *et al*., 2004, 2005). Dicho método de minimización es aplicable siempre que la función costo $\phi(\cdot)$ satisfaga la condición que $\exists\alpha_{max}/\forall\alpha\in[0;\alpha_{max}]$, $g_\alpha(x)= x^2/2-\alpha\varphi(x)$, es estrictamente convexa, que para el caso de la función costo de Huber se satisface completamente (Idier, 2001).

La minimización semi-cuadrática consiste en introducir un conjunto de variables auxiliares $\mathbf{d}= (d_1, d_2, \ldots, d_N)^T$ y a la redefinición de una función expandida K(**a, d**) la cual alcanza su mínimo en el mismo punto que la función J(·). Dicha función K(·,·)viene definida como:

$$K(\mathbf{a},\mathbf{d})= \frac{1}{\alpha}\sum_{i=1}^{N}\frac{1}{2}\left((y_i-(\mathbf{Ta})_i-\mathbf{d}_i)^2+\zeta_\alpha(\mathbf{d}_i)\right) \qquad (8)$$

donde la función ζ_α se define a partir de la función costo φ(·), como (Mazet *et al*., 2005)

$$\zeta_\alpha(d_i)= \sup_x\left(\alpha\varphi(x)-(x-d_i)^2/2\right) \qquad (9)$$

Se puede demostrar, que el nuevo criterio K(·,·)es cuadrático en **a** y convexo en **d**, justificando así el nombre de "criterio semi-cuadrático" (Mazet *et al*., 2004). La función Huber cumple con la condición anterior, para un valor de $\alpha_{max}= 0{,}5$. Así, la meta es entonces la minimización de la función expandida K(**a**, **d**).

Mazet *et al.* (2004, 2005) reportaron un algoritmo iterativo que permite resolver el problema de minimización dado por la Ec. 8. Dicho algoritmo, conocido como LEGEND, ha sido usado anteriormente por esos autores para estimar la línea base en espectros de infrarrojo y de Raman, y puede ser aplicado en este caso para la minimización de J(**a**). El algoritmo LEGEND estima iterativamente **a** y **d** y como sigue:

- Minimiza la función K(·, ·) dada por la Ec. 8 con respecto a **a**, dejando **d** fijo. Esto se reduce a la expresión

$$(\mathbf{T}^T\mathbf{T})\hat{\mathbf{a}} = \mathbf{T}^T(\mathbf{y}+\mathbf{d}), \Rightarrow$$
$$\hat{\mathbf{a}} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T(\mathbf{y}+\mathbf{d}) \quad (10)$$

resultado que puede interpretarse como la solución de los mínimos cuadrados cuando la señal es **y+d**.

- Dejando fijo **a** en la Ec. 8, el mínimo de la función K(·, ·) se alcanza si la i-enésima componente del vector **d** satisface

$$\forall i, \quad \hat{d}_i = -\varepsilon_i + \alpha\varphi'(\varepsilon_i) \quad (11)$$

donde $\varepsilon_i = y_i - b_i = y_i - (\mathbf{Ta})_i$ y φ' denota la derivada de la función costo.

La Figura 4 resume el algoritmo de LEGEND para la minimización de la función dada por la Ec. 8. Como se puede observar, el algoritmo iterativo se inicializa con la solución de mínimos cuadrados para una primera estimación de los parámetros del modelo y finaliza cuando la evolución relativa entre dos iteraciones sucesivas sea menor que un cierto valor umbral, U, pre-establecido. Para el caso en estudio la diferencia relativa de la función K(**a**, **d**) entre dos iteraciones sucesivas fue fijado en el orden de 10^{-5}.

De acuerdo a las pruebas realizadas por Mazet *et al.* (2004, 2005) y Mazet, (2005), se obtuvieron resultados apropiados con el uso del algoritmo LEGEND con el objetivo de corregir la LB en espectros infrarrojos y de Raman. En el presente caso, el desempeño del algoritmo de LEGEND fue pobre cuando se trató de aproximar la LB de datos electroforéticos usando un único polinomio, aún cuando su grado fuese elevado, debido fundamentalmente a la variabilidad dinámica y a la alta resolución que presentan los datos electroforéticos.

Como una posible alternativa a fin de superar estos inconvenientes, se propone un algoritmo que a diferencia del original, el cual resuelve un problema de optimización global usando todos los datos del registro electroforético, hace la corrección de la LB localmente. Es decir, se particiona el registro electroforético en ventanas no solapadas y se aplica el algoritmo LEGEND a cada ventana independientemente, obteniéndose así un mejor desempeño, en particular cuando se tienen picos de baja altura adyacentes a picos elevados, y que en este caso el especialista considera que la LB deba pasar por los mínimos locales del registro electroforético. Una muestra de los resultados conseguidos con el algoritmo de optimización aplicado localmente, se puede apreciar en la Figura 5. Obsérvese que la LB sigue el cambio producido por la presencia de picos de pequeña amplitud alrededor del pico elevado.

Dado que la optimización local se realiza independientemente en cada tramo del electroferograma (ventana de observación no solapada), se producen soluciones distintas en los bordes de las ventanas, ocasionando cambios abruptos que no forman parte de la LB real y que se deben a la concatenación de las soluciones parciales (locales) para conformar la solución total. A fin de superar esta limitación es necesario suavizar la curva obtenida y así suprimir estos cambios abruptos en los puntos de unión del final de una subsección con el comienzo de la siguiente. Para la suavización de la curva final de la señal LB, se utilizó el filtro suavizador de Savitzky y Golay (1964), el cual realiza una regresión de tipo polinomial local para estimar los puntos de la curva suavizada.

Paso 1: Inicializar:

$$k = 0$$
$$\hat{\mathbf{a}}^{(0)} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{y}$$
$$\forall i, \quad \hat{d}_i^{(0)} = 0$$

Paso 2: Determinar las variables auxiliares:

$$\forall i, \quad \hat{d}_i^{(k+1)} = -\varepsilon_i + \alpha\varphi'(\varepsilon_i)$$

Paso 3: Estimar los nuevos parámetros de la línea base

$$\hat{\mathbf{a}}^{(k+1)} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T(\mathbf{y}+\mathbf{d}^{(k+1)})$$

Paso 4: Comprobación de convergencia:

Si $\left|K(\hat{\mathbf{a}}^{(k+1)}, \hat{\mathbf{d}}^{(k+1)}) - K(\hat{\mathbf{a}}^{(k)}, \hat{\mathbf{d}}^{(k)})\right| > U$, $k = k+1$

ir al Paso 2, de lo contrario finalizar.

Figura 4. Resumen del algoritmo LEGEND.

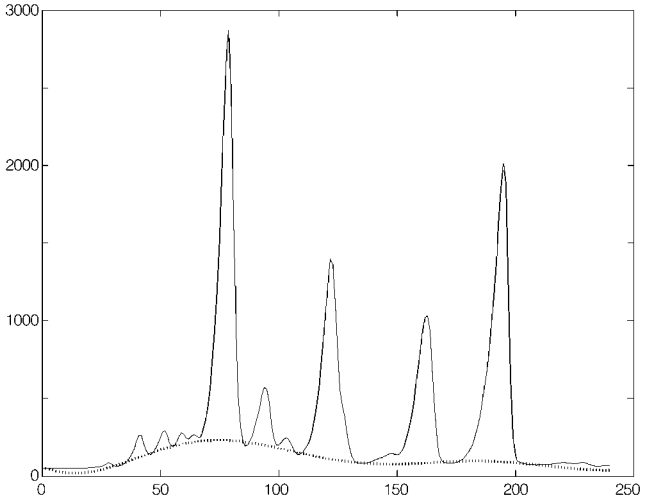

Figura 5. Línea base conseguida con LEGEND optimizado localmente.

Evaluación del Algoritmo Propuesto

A fin de evaluar el algoritmo desarrollado, éste se aplicó a una base de datos de registros electroforéticos, los cuales fueron tomados en un equipo de electroforesis capilar desarrollado en el Laboratorio de Fisiología de la Conducta, Universidad de Los Andes, Venezuela. Los registros electroforéticos seleccionados como datos de prueba fueron pre-acondicionados como sigue.

Se aplicó la transformada *wavelet* discreta hasta el sexto nivel de descomposición. Seguidamente, operaciones de umbralización dada por la Ec. 4 son aplicados a los coeficientes de detalles en los niveles 5 y 6. La señal es luego reconstruida hasta el cuarto nivel de descomposición, obteniéndose así una señal a baja resolución con componentes de ruido suprimidas. El valor del umbral escogido fue fijado a $u = C_1\sqrt{2\log(N_j)}$, donde N_j: número de coeficientes *wavelet* de detalles al j-enésimo nivel de descomposición. Se utilizó la *wavelet symlet 4* según las recomendaciones de Weidong *et al.* (2003) y se escogieron los niveles de descomposición y el valor del umbral según Ceballos *et al.* (2008).

En la etapa de estimación de LB, los criterios seguidos para definir el tamaño de la ventana de observación, el grado del polinomio que modela la LB así como el valor del umbral de la función costo fueron 1) conseguir una estimación de la LB apropiada cuando existen picos grandes que arrastran a los picos pequeños presentes en las adyacencias, y 2) preservar en todo momento de los picos pequeños asociados a bajas concentraciones. Se probó el algoritmo propuesto con dos diferentes tamaños de ventanas no solapadas, N/3 y N/5, siendo N el número de muestras de la señal electroforética pre-acondicionada.

Para el caso de los datos fraccionados en tres partes se obtuvo un mejor desempeño con un polinomio

de grado 9 y con un valor umbral de la función costo fijado en 0,8; mientras que para el caso en que el registro electroforético se fraccionó en cinco partes, el grado del polinomio fue 5 y el valor del umbral fue fijado en 2.

Para propósitos de comparación, se implementó el algoritmo desarrollado por Ceballos *et al*. (2007, 2008) a fin de contrastar ambos métodos. Una muestra de los resultados se aprecia en la Figura 6, donde se presentan las LB obtenidas con el algoritmo desarrollado y la LB obtenida con el algoritmo propuesto por Ceballos *et al*., que se denotará en lo sucesivo como método no paramétrico dado que no utiliza un modelo parametrizado de la LB y por lo tanto no se requiere un proceso de optimización. Como se puede observar en la Figura 6 el método no paramétrico falla en seguir el comportamiento dinámico de la LB en los alrededores de picos elevados (instante de migración 300).

Adicionalmente, al comparar el desempeño del algoritmo para distintos tamaños de ventana, cabe destacar el hecho de que al disminuir el tamaño de la ventana el orden del polinomio se reduce y la curva obtenida como LB trata de ajustarse aún más a los mínimos locales del electroferograma. A fin de observar estas diferencias la Figura 6 también muestra las líneas base obtenidas con los datos fraccionados en tres y cinco partes. En el instante de migración 150 se nota la marcada diferencia entre ambas líneas base, pudiéndose apreciar como la línea base conseguida con los datos seccionados en cinco partes es la que trata de buscar el mínimo local presente en el instante 142.

Para el caso de picos pequeños, asociados con sustancias de baja concentración que pudieran ser de interés para el especialista, los mismos se preservan cuando se corrige la LB usando el método propuesto, mientras que para el método no paramétrico esto no siempre ocurre, tal como se puede apreciar en la Figura 7, donde se observa un tramo de un electroferograma corregido. El método no paramétrico elimina por completo los picos existentes en la señal original en los instantes de migración entre 490 y 550. Así mismo, los picos ubicados en 575 y 715 son suprimidos casi por completo.

Dado que la inspección visual pudiera ser subjetiva y a fin de validar numéricamente el algoritmo desarrollado, se probaron los algoritmos en una aplicación experimental, donde se desea determinar la cantidad de glutamato en muestras tomadas en dos diferentes áreas del cerebro en 24 ratas (Ceballos *et al*., 2008). El valor de la concentración de glutamato tomado del electroferograma acondicionado (preprocesamiento *wavelet* seguido de corrección de LB) se comparó con respecto al valor obtenido, corrigiéndose manualmente la línea base en la zona donde el glutamato aparece en el electroferograma. Este valor, medido por el especialista, es la referencia de comparación y es considerado como el valor verdadero de la concentración de esta sustancia en la muestra en estudio. Se determinó el error relativo como $\varepsilon_r = (V_{est}-V_{esp})/V_{esp}\times 100$, donde V_{est} representa el valor medido luego de realizar la corrección de LB con el algoritmo propuesto y V_{esp} es el valor medido por un especialista luego de una corrección manual de la línea base.

Para este grupo de registros electroforéticos el grado del polinomio para los datos seccionados en 3 partes fue de 12 y el umbral k= 2, y para el caso de los datos particionados en 5 el grado del polinomio fue 6 y el umbral fue 1,4. En la Figura 8 se muestra el error relativo después de corregir la LB usando los distintos métodos. Se puede apreciar como el método propuesto tiene en general un mejor desempeño en comparación con el método no paramétrico, consiguiéndose un error promedio de 1,45% para 3 partes y 1,58% para cinco partes en comparación con 2,36% para el método no paramétrico. Similarmente, la desviación estándar del error para cada caso es 1,38, 1,58 y 2,02 respectivamente.

En cuanto a la complejidad del algoritmo usando como figura de mérito el tiempo de cómputo tomado por un computador Pentium IV, de 3GHz y 512MB de RAM, el tiempo de ejecución del algoritmo propuesto es de 6s en promedio para cada señal pre-acondicionada de ~750 muestras.

Conclusiones

Se ha desarrollado un método para estimar la línea base (LB) de las señales electroforéticas, mediante la aplicación del algoritmo de optimización LEGEND en forma local en el dominio *wavelet*. Se constató que una función polinomial puede modelar adecuadamente la señal de LB presente en los electro-

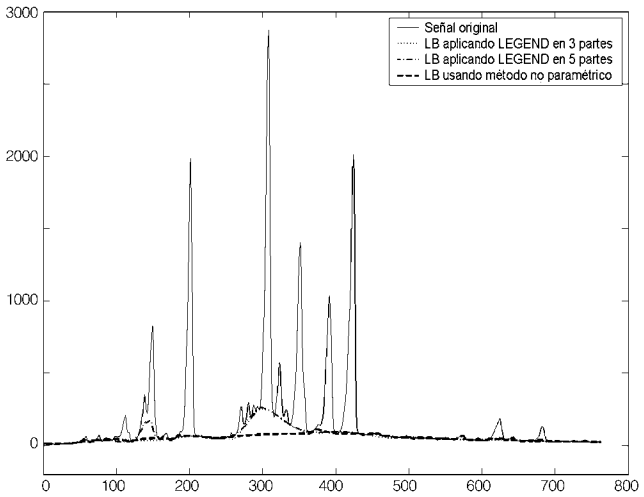

Figura 6. Comparación de líneas bases entre LEGEND y el método no paramétrico.

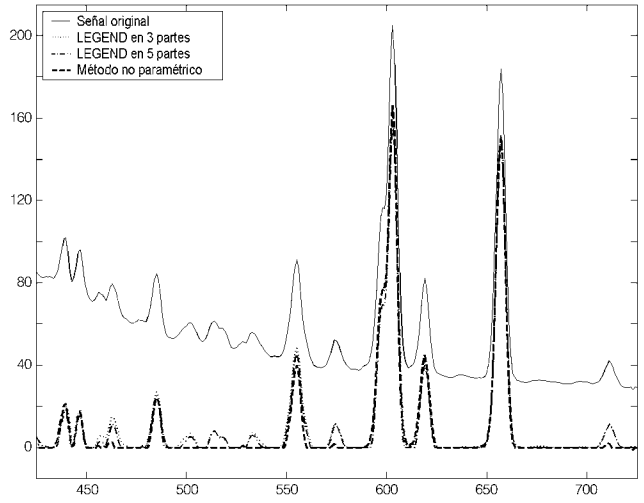

Figura 7. Tramo de un electroferograma corregido.

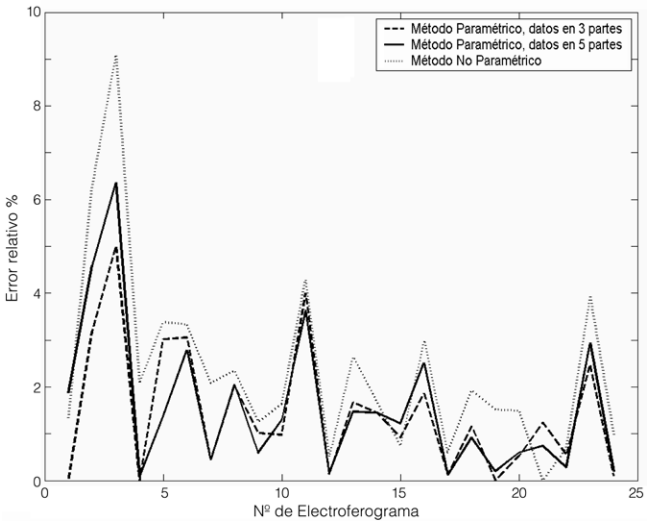

Figura 8. Errores conseguidos con cada método en la medición de la cantidad de glutamato.

ferogramas usando funciones de costo asimétricas robustas para la estimación de los parámetros del modelo polinomial. El algoritmo propuesto corrige efectivamente la LB, en particular en las zonas donde se aprecian picos de baja altura en las adyacencias de picos elevados. Adicionalmente, se preservan picos pequeños que representan información de baja concentración de algunas componentes químicas de la muestra en estudio.

AGRADECIMIENTOS

Los autores agradecen a Luis Hernández y Gerardo Ceballos por las discusiones que contribuyeron a este trabajo, el cual ha sido financiada por el Fondo Nacional de Ciencia, Tecnología e Innovación (FONACIT, Proyecto G-2005000342) y por el Consejo de Desarrollo Científico, Humanístico y Tecnológico de la Universidad de Los Andes (CDCHT-ULA), bajo el código I-1043-07-02-F.

REFERENCIAS

Alexandra S, Francia F, Morela F (2003) Caracterización de clones de yuca (manihot esculenta) mediante marcadores protéicos e isoenzimáticos. *Interciencia 28*: 690-698.

Ceballos G, Paredes JL, Hernández L (2007) A novel approach for pattern recognition in capillary electrophoresis data. *Proc. IV Lat. Am. Cong. Biomed. Eng. 2007 Bioengineering Solutions for Latin America Health*. pp. 150-153.

Ceballos G, Paredes JL, Hernández L (2008) Pattern recognition in capillary electrophoresis data using dynamic programming in the wavelet domain. *Electrophoresis 29*: 2828-2840.

Donoho DL (1995) De-noising by soft-thresholding. *IEEE Trans. Inf. Theory 41*: 613-627.

Gamero MAR (2001) *Desarrollo de Nuevas Metodologías Analíticas en el Control de Calidad de la Industria Farmacéutica*. Tesis. Universidad Autónoma de Barcelona. España. 26 pp.

Huber PJ (2004) *Robust Statistics*. Series in Probability and Mathematical Statistics. Wiley. Nueva York, NY, EEUU. 308 pp.

Idier J (2001) Convex half-quadratic criteria and interacting auxiliary variables for image restoration. *IEEE Trans. Image Proc. 10*: 1001-1009.

Mazet V (2005) *Développement de Méthodes de Traitement de Signaux Spectroscopiques: Estimation de la Ligne de Base et du Spectre de Raies*. Tesis. Université Henri Poincaré. Nancy, Francia. 156 pp.

Mazet V, Brie D, Idier J (2004) Baseline spectrum estimation using half-quadratic minimization. *EUSIPCO*. Viena, Austria. pp. 305-308.

Mazet V, Carteret C, Brie D, Idier J, Humbert B (2005) Background removal from spectra by designing and minimizing a non-quadratic cost function. *Chemometr. Intell. Lab. Syst. 76*: 121-133.

Perrin C, Walczak B, Massart D (2001) The use of wavelets for signal denoising in capillary electrophoresis. *Anal. Chem. 73*: 4903-4917.

Savitzky A, Golay MJ (1964) Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem. 36*: 1627-1639.

Shadle SE, Allen DF, Guo H, Pogozelski WK, Bashkin JS, Tullius TD (1997) Quantitative analysis of electrophoresis data: novel curve fitting methodology and its application to the determination of a protein-DNA binding constant. *Nucl. Ac. Res. 25*: 850-860.

Sidney BC, Ramesh AG, Haitao G (1998) *Introduction to Wavelets and Wavelet Transform*. Prentice Hall. Upper Saddle River, NJ, EEUU. 256 pp.

Weidong C, Xiaoyan C, Xiurong Y, Erkang W, (2003) "Discrete wavelets transform for signal denoising in capillary electrophoresis with electrochemiluminescence detection," *Electrophoresis 24*: 3124-3130.

Weiping Y, Chengcai X, Jianhua L, Huang J (2006) The use of discrete wavelets for signal denoising in microchip capillary electrophoresis. *Proc. 6th World Cong. on Intelligent Control and Automation*. WCICA'2006. Vol. 2. pp. 5254-5258.

A BASELINE CORRECTION ALGORITHM FOR CAPILLARY ELECTROPHORESIS DATA USING LOCAL OPTIMIZATION OF THE LEGEND ALGORITHM IN THE WAVELET DOMAIN

José L. Paredes and Enedina Sosa

SUMMARY

A baseline (BL) correction algorithm for capillary electrophoresis (CE) data is developed. The proposed algorithm exploits a low-resolution wavelet representation of the original signal to locally model the BL as a polinomial function and applies the LEGEND optimization algorithm to obtain the model parameters, such that a robust non-symmetric cost function is minimized. The proposed algorithm outputs a suitable BL correction on those subsections of the electropherogram where peaks related to low concentration substances are near those of high concentration substances. The performance of the proposed algorithm was tested by measuring the glutamate mass on 24 electropherograms after the BL had been suppressed using the proposed algorithm. The resulting values are compared to those yielded by a conventional measuring method performed by a CE specialist and by a second CE baseline correction method recently introduced.

CORREÇÃO DE LINHA BASE EM DADOS ELECTROFORÉTICOS USANDO OTIMIZAÇÃO LOCAL DO ALGORITMO "LEGEND" NO DOMÍNIO WAVELET

José L. Paredes e Enedina Sosa

RESUMO

É proposto um método de correção da linha base (LB) de sinais eletroforéticos que explora a representação wavelet a baixa resolução do sinal original. A LB é modelada no dominio wavelet como uma função polinomial e se usa o algoritmo de otimização LEGEND para determinar os parâmetros do modelo polinomial que melhor se ajusta a uma subseção da LB, de forma tal de minimizar uma função custo assimétrica robusta. O algoritmo proposto produz uma correção adequada da LB naquelas áreas do eletroferograma onde se apreciam substâncias de baixa concentração nas adjacências de substâncias de concentração elevada, preservando picos associados com as primeiras. Validou-se o algoritmo desenvolvido em um problema de medição da quantidade de glutamato presente em 24 registros eletroforéticos e foi comparado seu desempenho com os valores medidos pelo especialista onde a linha base é corrigida em forma manual e com os valores emitidos por um segundo algoritmo de correção de LB recentemente proposto.