ELECTROPHORETIC REGISTERS: CORRECCIÓN DE LÍNEA BASE, ALINEAMIENTO, ELIMINACIÓN DE RUIDO Y CONCENTRACION DE ENERGÍA

L. E. Mendoza[1]., Leonor Jaimes C[2]., H.J. Velandia[3]

[1,3] Universidad de Pamplona. Grupo de Investigación GIBUP. Programa de Ingeniería en Telecomunicaciones.
[2] Universidad de Pamplona. Grupo de investigación INGAPO. Programa de Ingeniería Industrial.

Resumen: Este artículo presenta una nueva metodología para la corrección de línea base, el alineamiento de picos, la eliminación de ruido y la clasificación de datos electroforéticos (**DER**). Las técnicas matemáticas usadas fueron: corrección linea base L2, análisis wavelet, alineamiento Nedleman and Wusch y análisis multi energía que fue utilizada con el propósito de ubicar información importante en los aminoácidos presentes en los datos electroforéticos. Adicionalmente, este artículo evidencia que es posible corregir la línea base, eliminar ruido y alinear datos electroforéticos sin cambiar sus características originales de energía y formas de onda. Se utilizaron 35 señales, en las cuales el porcentaje de error en cuanto a corrección de línea base fue de 0%, en alineamiento fue de 1.23%, y de clasificación fue de 3.5±1.2. Finalmente, se muestra que es posible utilizar esta metodología sin importar las características de las señales originales, lo que permite que sea un sistema de múltiples soluciones en los datos electroforéticos, cromatográficos y otros.

Palabras clave: datos electroforéticos, Clasificación, corrección línea base, alineamientos de picos, procesamiento de señales.

ELECTROPHORETIC REGISTERS: CORRECTING BASELINE, ALIGNMENT, DENOISING AND ENERGY CONCENTRATION

Abstract: This paper presents a methodology new for baseline correction, peak alignment, denoising and electrophoretic data (**DER**) classification. The mathematical techniques used were based line correction L2, wavelet analysis, alignment and Wusch Nedleman and multi energy, that was used for the purpose of locating important information on the amino acids present in electrophoretic data. Additionally, this article evidence that it is possible to correct the baseline, eliminate noise and align electrophoretic data without changing their original energy characteristics and waveforms. 50 signals were used, in which the error rate in terms of baseline correction was 0%, in alignment was 1.23%, and classification was 3.5 ± 1.2. Finally, we show that this method can be used regardless of the characteristics of the original signals, which allows it a system of multiple solutions in the electrophoretic, chromatographic and other data.

Keywords: electrophoretic data, classification, baseline correction, peak alignment, signal processing.

***Para citar este artículo:** Mendoza L.E., Jaimes C L., Velandia.H.J. Electrophoretic regosters:Corrección de línea base, alineamiento,eliminación de ruido y concentración de energía.Revista Bistua.2015.13(2):3-11

+ Autor para el envió de correspondencia y la solicitud de las separatas: Mendoza LE. Universidad de Pamplona. Programa Ingeniería en Telecomunicaciones.. Facultad de Ingenierías y Arquitectura. email: luis.mendoza@unipamplona.edu.co

Recibido: Octubre 08 de 2014 Aceptado: Marzo 10 de 2015

1. INTRODUCCIÓN

La Electroforesis Capilar es una técnica de separación de los diferentes compuestos químicos que conforman una muestra específica. Esta técnica es muy utilizada en la actualidad en diferentes áreas de investigación, como la Bioquímica, ingeniería, Microbiología, Biología celular, Biotecnología, Medicina Forense e Industrias Alimentarías entre otras [1,2,3], pues a partir de los registros adquiridos es posible identificar y posteriormente medir la concentración de las distintas sustancias que conforman la muestra en estudio. Una aplicación de particular interés de la electroforesis capilar es el análisis de muestras tomadas en individuos que se sospecha de tener el virus de la meningitis, con el objetivo de lograr la detección temprana de esta enfermedad y así prevenir a futuro daños cerebrales, epilepsias, hipoacusia o sordera, hidrocefalia, pérdida de la visión o incluso hasta la muerte en casos extremos[4,5].

La enfermedad de meningitis también conocida bajo el nombre meningitis espinal, es una infección de los fluidos de la médula espinal y de los fluidos que rodean al cerebro [6,7]. Esta enfermedad puede ser causada por una bacteria o por una infección viral, en este trabajo los individuos en estudio fueron contaminados con la bacteria "*klebsiella pneumoniae*", es posible que estos síntomas no aparezcan o no se detecten de forma asintopatica. De ahí surge la necesidad de desarrollar algoritmos que a partir de registro electroforéticos sea posible detectar el virus en sus distintas etapas de su desarrollo.

Sin embargo, en el análisis de datos electroforéticos provenientes de muestras tomadas de pacientes que se sospecha tener el virus de la meningitis, diferentes inconvenientes hay que enfrentar cuando se quiere realizar un estudio que involucre la corrección de línea base, el alineamiento la eliminación de ruido y el reconocimiento de patrones para procesos de clasificación. Problemas como la

poca reproducibilidad de los datos y superposición de picos, son algunos de los inconvenientes que debe enfrentar el investigador. Los datos electroforéticos en muchas ocasiones son analizados de forma visual con el fin de evitar errores de análisis debidos a los problemas anteriormente mencionados. Sin embargo, si se requiere un análisis de numerosas muestras asociadas a diferentes individuos, sería imposible encontrar un buen resultado y en un tiempo relativamente corto si el proceso de análisis se hace en la forma tradicional, que involucra un análisis asistido de los registros electroforéticos en búsqueda de patrones específicos asociados a la presencia o no del virus. Este análisis manual induciría a errores debido al agotamiento físico del especialista. Por este motivo se hace necesario el desarrollo de herramientas auxiliares basadas en un computador y soportadas matemáticamente para la selección y extracción de características relevantes que le indiquen al especialista en forma rápida y precisa la existencia o no del virus de la meningitis. En la actualidad diferentes trabajos de investigación [8,9,10] han utilizado técnicas matemáticas aplicadas a datos electroforéticos con el objetivo de extraer, seleccionar y clasificar diferentes anormalidades que pueden presentar ciertas sustancias de interés, así como también la detección de ciertas patologías[9,10,11], logrando la caracterización o parametrización de las diferencias o similitudes encontradas en los registros electroforéticos. Otro trabajo que ha aportado un avance significativo en el procesamiento de datos electroforéticos usando análisis wavelet y programación dinámica con el fin de acondicionar los datos electroforéticos y hacer una reconocimiento de patrones en mejores condiciones fue recientemente introducido en[12].

El aporte de este trabajo es el alineamiento de picos, corrección de línea base y reconocimiento de patrones presentes en datos electroforéticos que permitan identificar zonas importantes, en base al estudio de características presentes en registros electroforéticos, con ayuda de técnicas como: alineamiento, coreccion de línea base, eliminación de ruido y análisis de energía. En un futuro, el uso de herramientas matemáticas de clasificación no lineales como son las máquinas de soporte vectorial tradicionales (SVM) y las máquinas de soporte vectorial de mínimos cuadrados (LS-SVM) recientemente introducidas[13,14,15], podrían lograr una discriminación lo suficientemente acorde para realizar la clasificación de datos electroforéticos contaminados con el virus de la meningitis y así poder detectar en fases tempranas (de 24 -72 horas) el patrón que identifique la existencia o no del virus.

2. CORRECCIÓN DE LINEA BASE, L_2

En este artículo se muestra el proceso matemático, que se utiliza para corregir el problema de línea base, que tiene todos los datos electroforéticos, cromatográficas entre otros. Es importante mencionar que esta aplicación fue realizada por el grupo de investigación GIBUP. Por notación **X** es la señal cromatrográfica, de longitud N. Entonces, en Ecu 1, se muestra como es el proceso para encontrar los valles de la señal.

$$\mathbf{X}(i-1) < \mathbf{X}(i) \;\&\; \mathbf{X}(i+1) < \mathbf{X}(i) \qquad (1)$$

Una vez se hallan los valles presentes en la señal, se busca la ecuación de la recta que satisface la unión del primera valle al segundo, del segundo al tercero y así sucesivamente. La Ecu 2, muestra la estructura que satisface las rectas que unen los valles presentes en la señal.

$$Y_{i,k}=M_i\, x+ b_i \quad (2)$$

Donde i=**v**-1 y **v** son el número de valles y k está definido como la longitud de datos presentes en cada recta i y tiene la misma longitud X. Seguidamente al encontrar todas ecuaciones que unen los valles, se procede a buscar las imágenes de los valores de la variable independiente entre cada recta creada. Esto con el fin de encontrar el área bajo la curva real que tiene cada pico, sin importante si las líneas bases de los picos presentes en la señal están con valores no iguales. La ecuación que modela la corrección de línea base final está dada en Ecu 3. Esta estructura matemática permite tener la corrección de línea base de cualquier señal electroforética o cromotográfica.

$$G_k=Y_{i,k}-X_k \quad (3)$$

3. WAVELET Y ALINEAMIENTO DE PICOS

Sea $f(t)=x(t)+e(t)$ el modelo matemático de la señal electroforética obtenida del proceso de separación de sustancias, donde $e(t)$ se define como el ruido introducido en todo el proceso de adquisición y $x(t)$ corresponde a la señal de interés, registro electroforético. En general el ruido $e(t)$ es debido a diferentes factores tales como: cambios de temperatura, cambios de voltaje, detector no calibrado o capilar en malas condiciones, entre otras[16,17]. El objetivo en este contexto es encontrar un estimado de la señal $x(t)$ que conserve las características más importantes de la señal, eliminando las componentes de ruido de la misma, la región de actividad nula y los cambios abruptos que pudiera presentar la señal $e(t)$. En este trabajo se usó la transformada wavelet como herramienta para estimar la señal $x(t)$, suprimiendo las componentes de ruido sin alterar considerablemente las componentes de interés. Otras aplicaciones que se han desarrollado en trabajos recientes haciendo uso de la TW fueron: reconocimiento de zonas de interés, disminución de resolución de la señal electroforética, y estimación de la línea base.

La **transformada wavelet** continua se define matemáticamente como:

$$C_{x,y}=\frac{1}{\sqrt{y}}\int_{-\infty}^{\infty} f(t)*\psi\left(\frac{t-x}{y}\right)dt \quad (4)$$

Donde $f(t)$ es la señal en estudio (señal electroforética) y $\psi(t)$ es conocida como la función wavelet madre. El parámetro x es la ubicación en el tiempo de la función $\psi(t)$ mientras que el parámetro y está asociado a la escala o el ancho de la función $\psi(t)$. En (2.1), $C_{x,y}$ se denomina coeficiente wavelet y representa el grado de similitud de una parte de la señal con la señal $\psi(t)$. Al aplicar la TWC a una señal se obtiene como resultado un arreglo bidimensional donde las variables escala (y) y traslación (x) toman valores en el conjunto de los números reales. En este trabajo se

usó análisis wavelet, para eliminación de ruido en datos electroforéticos con el fin de aplicar de manera más efectiva la corrección de línea base ya que la señal original viene en su estado original con alto grado de variabilidad en sus zonas de no interés.

Dos métodos son usados con el objetivo de modificar los coeficientes mediante el proceso de umbralización: umbralización rígida (*hard*) y umbralización suave (*soft*). Estos métodos son matemáticamente descritos en (5) y (6) respectivamente.

$$\hat{d}_{j,k} = \begin{cases} d_{j,k} & si \quad |d_{j,k}| > Th \\ 0 & si \quad |d_{j,k}| \le Th \end{cases} \tag{5}$$

$$\hat{d}_{j,k} = \begin{cases} sig(d_{j,k}) * (d_{j,k} - Th) & si \quad |d_{j,k}| > Th \\ 0 & si \quad |d_{j,k}| \le Th \end{cases} \tag{6}$$

Aquí $\hat{d}_{j,k}$ representa los coeficientes de detalles del nivel de descomposición j después de ser umbralizado, Th el umbral elegido y sig(x) la función signo definida como: +1 si x≥0 y -1 si x<0. Nótese que si se utiliza una umbralización rígida los coeficientes que se preservan dejándolos inalterados son aquellos cuyo valor absoluto supera el umbral, mientras para otro caso el coeficiente toma el valor de cero.

Por otro lado la programación dinámica permite realizar un alineamiento de picos de manera automática: Este algoritmo también se conoce con el nombre de Nedleman & Wunsch[18].

Considere la matriz de recursividad S, también denominada matriz de programación dinámica, formada por n filas y m columnas definida como:

Condiciones iniciales: Matriz de llenado

$$\begin{aligned} &S(0,0) = 0 \\ &S(i,0) = id \quad 1 \le i \le n \\ &S(0,j) = jd \quad 1 \le j \le m \end{aligned} \tag{7}$$

$$S(i,j) = \max \begin{cases} S(i-1, j-1) + Score(x_i, y_i) \\ S(i, j-1) + d \\ S(i-1, j) + d \end{cases}$$

Donde d es un valor entero (real) que establece la penalización por inserción de espacios[18] y la función $Score(x_i, y_i)$ define el valor de las ponderaciones otorgado al alinear el elemento i de la secuencia x con el elemento j de la secuencia y.

Nótese que el valor a ser asignado a la i,j-enésima componente de la matriz S depende solo de las posiciones izquierda, superior y diagonal a $S(i,j)$ y de la penalización d y el valor *Score* $S(i,j)$. Obsérvese que además el valor que toma $S(i,j)$ en cada posición depende del valor máximo generado por cualquiera de las tres posiciones (izquierda, superior y diagonal), la correspondiente penalización por inserción de espacios S. Dicho de otra manera, la posición que genere el mayor valor será la elegida para definir el valor en $S(i,j)$. El valor que toma $S(i,j)$ en dicha posición, es el valor máximo generado por (2.14). Teniendo en cuenta lo mencionado anteriormente la matriz $S(i,j)$ se llena hasta su máxima posición $S(n,m)$.

Una vez que se tiene la matriz $S(i,j)$ totalmente llena, se procede a la reconstrucción del camino óptimo (alineamiento óptimo). El movimiento que se elige para encontrar el camino óptimo dependerá de la posición que generó el valor que toma la casilla

de la matriz $S(i, j)$ en ese momento. Se debe tener en cuenta que el alineamiento global se comienza en la última columna hasta encontrar $S(0,0)$.

4. CONCENTRACIÓN DE ENERGÍA

El proceso de clasificación usando concentración se realizó con el objetivo de cuantificar el error de los procesos anteriormente mencionados y así mismo presentar resultados de los aminoácidos que en cierto momento presentaron cambios de energía. Matemáticamente la energía se describe en la ecu 8.

$$E = \text{sum}(X_i \hat{} 2) \quad (8)$$

Esta representación busca en este trabajo identificar los valores de energía que contiene cada aminoácido en el dato electroforético, esto permitirá en un futuro realizar análisis cuantitativos en cambios de energía en diferentes picos del mismo aminoácido.

Finalmente, la fig. 1, muestra un ejemplo tipo de un electroferograma, donde se videncia 3 zonas, la zona 1 y 3 son definidas como zonas de no actividad y la zona 2 es la zona donde se tiene la información importante en la señal. Esta señal es típico en las 50 señales donde se realizaron pruebas para verificar la eficiencia del algoritmo.

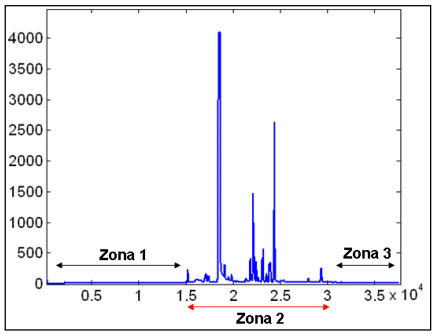

Figura 1 - Ejemplo DER.

4. RESULTADOS

La corrección de línea base es un procedimiento importante en el análisis de la energía y de las características de cada uno de aminoácidos presente en la muestra, como no están sobre la misma línea base, existen sistema que no tiene en cuenta esta situación y por ende sus análisis son erróneos. La fig. 2, muestra los resultados de la ubicación de las rectas que unen cada uno de los picos con el fin de realizar la corrección de línea base. Vale la pena mencionar que el sistema tiene 0% de error, esto se calculó por medio de la energía de cada uno de los aminoácidos, presentes en la muestra.

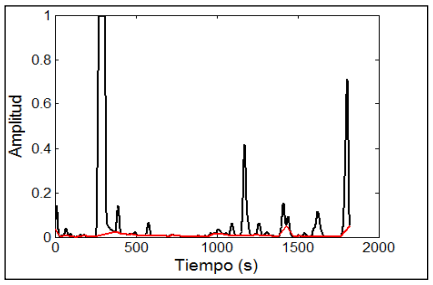

Figura 2 – Ubicación rectas para corrección línea base.

A continuación se muestra (ver fig. 3) la corrección de línea base de la señal electroforética, en las posiciones de señal de 400 a 1100 puntos en el tiempo, nótese como es evidencia que los valles de cada pico fueron bajados a una línea base 0.

Seguidamente se muestran los resultados de alineamiento de secuencias. En la fig. 4 se observa el resultado de un alineamiento de datos electroforéticos. Nótese, como inicialmente están separados los picos, y una vez aplicado el algoritmo de alineamiento, los picos de las mismas clases quedan superpuestos, este algoritmo se realizó en 50 señales.

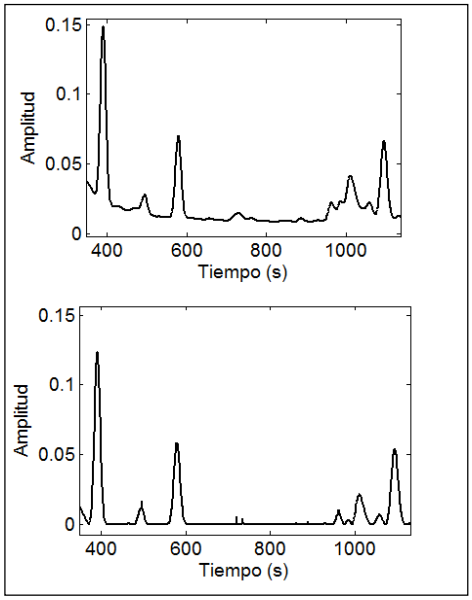

Figura 3 – Corrección línea base.

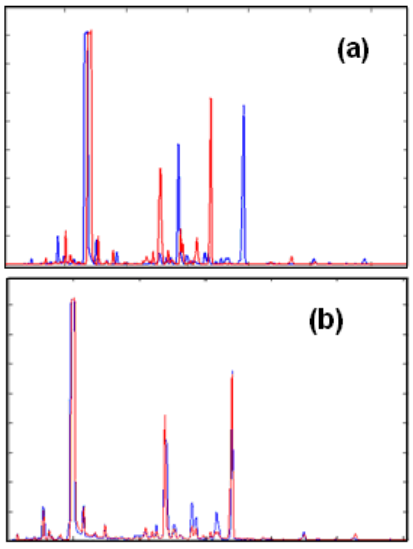

Figure 4 - DRE, a) registro no alineado y b) registro alineado.

La eliminación de artefactos poco relevantes se realiza con el propósito de mejorar las condiciones de la señal en los procesos de corrección línea base y alineamiento. En la fig. 5, se muestran los resultados de aplicar un umbral hard usando transformada wavelet. Se usó una wavelet madre db5 y un nivel de descomposición 3. A cada coeficiente de detalle se aplicó el filtrado hard. Se usó un umbral universal wavelet. En el círculo marcado de color rojo se puede observar que han sido removidos artefactos que son pocos relevantes y que hacen que los cálculos anteriormente mencionados no sean los esperados.

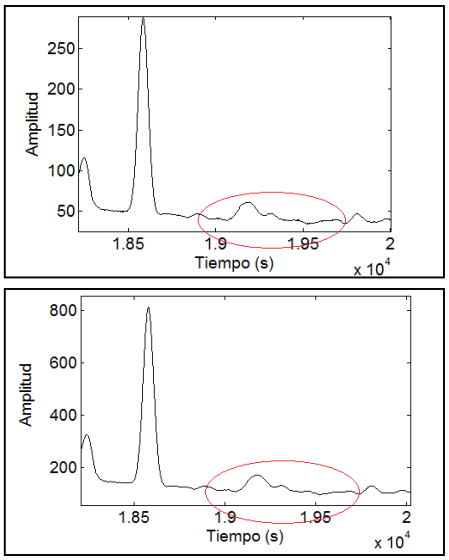

Figura 5 – Eliminación de ruido en DER.

Finalmente, se muestran los resultados de los cálculos de energía de cada uno de los picos (ver fig. 6), de la señal original, con corrección de línea base, y con alineamiento, esto con el fin de demostrar que no se alteraron los cálculos de energía durante los proceso. Nótese como las concentraciones de energía son idénticas para la señal original (datos color negro) como para la corrección de línea base (datos color rojos).

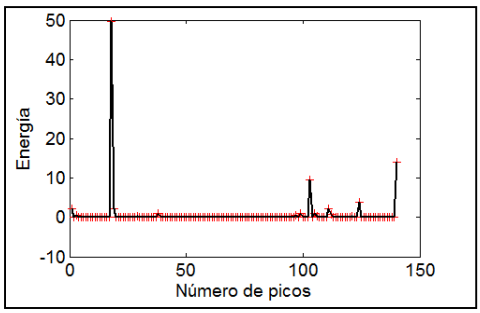

Figura 6 – Calculo de energía de cada pico.

Es importante aclarar que el análisis wavelet solo se utilizó para mejor condiciones y lograr ubicar donde se encontraban los puntos que debían ir a cero en la señal para la corrección de línea base, aquí no se hizo análisis de energía porque es claro que las energía cambia del espacio del tiempo al espacio de la escala frecuencia que es el espacio wavelet. Finalmente se muestran los resultados de energía una vez aplicado el alineamiento de secuencias, el error en este caso fue de 3.5±1.2, esto ya que realizaron inserciones de datos, que permitían correr un pico de su posición original, aquí existe una discusión en el tema de tiempo de retención por ejemplo en cromatografía, pero es importante aclara que se hace el alineamiento ya que en muchas ocasiones los tiempo de retención cambian y esto permite que los análisis puntuales de compuesto o aminoácidos no se haga de manera objetiva, es así como una vez realizado el alineamiento, se procedió a encontrar las energía para cada uno de los picos y al final se obtuvo una señal idéntica a las Fig. 6. Lo que permite evidenciar que la energía en cada uno de los picos es igual a las energías originales.

5. CONCLUSIONES

Se logró demostrar que es posible corregir línea base y alinear picos en datos electroforéticos sin realizar ninguna medicación en concentración de energía de los N picos de cada señal. Por otro lado el procedimiento desarrollado permite realizar análisis visuales mucho más exactos, ya que se tienen los datos en un mismo rango y con alineaciones entre los mismos aminoácidos en señales diferentes. El sistema se probó en más de 50 señales, con el fin de revisar el funcionamiento y aplicabilidad generalizada con todas las variables que tiene cada señal y se logró demostrar que sin importar el comportamiento de la señal, el sistema funciona correctamente y siempre mantienen las mismas concentraciones de energías iguales a los valores originales, esto indica que no se modifican las características principales de los datos como lo son, energía y forma del registro.

En análisis wavelet solo se usó para disminuir el rizado en las señales y esto generó un espacio de trabajo mucho más eficiente y fue gracias al uso de la transformada wavelet estacionaria y su nivel de descomposición 3. En cuanto al alineamiento es importante resaltar que los picos no se modifican, lo que se modifican son las zonas donde no existe picos, la inserción de datos iguales a cero generó resultados importantes ya que realizar los corrimientos de los picos necesarios sin modificar las concentraciones de energía del electroferograma general y menos aún la energía de los picos por separado. Finalmente este sistema permite a los expertos tener una herramienta más generalizada de las que actualmente existen y con posibilidades de ingresar mayores desarrollos en un futuro.

Comentarios finales donde se resumen y se puntualiza sobre los

aportes más significativos del trabajo.

REFERENCIAS

[1] Berruela L, R. Alonso R, K. Héberger K. Supervised pattern recognition in food analysis. Chromatrography, Elsevier, 2007:196-214.

[2] Lanz C, U, Thormann W. Capillary zone electrophoresis with a dynamic double coating for analysis of carbohydrate-deficient transferring in human serum precision performance and pattern recognition. Chromatrography, Elsevier, 2003:131-147.

[3] La S, Cho J, Han Kim J, K. Rao K. Capillary electrophoretic profiling and pattern recognition análysis of urinary nucleosides from thyroid cancer patients. Analytica chimica. 2003:171-182.

[4] Michal M, Smiatacz T, Hlebowicz M,Pajuro R & H. Trocha H**.** Coagulation, and outcome in bacterial meningitis an observational study of 38 adults cases. Cromatography, 2008.

[5] Misal J, Embon A, Darawshe A, Kidon M & and Magen E. Community acquired acute bacterial meningitis in children and adults: An 11 year survery in a community hospital in Israel, meningitis, 2008.

[6] Gong Z, Fena E, Zhang Q & Xiu Z. Computational method for inferring objective function of glycerol metabolism in Klebsiella pneumoniae. Computational Biology and Chemistry. 2008.

[7] Pasa S, Altintas A, Cil T, Ustun C, Bayan K, Danis R, Urakci Z, Tuzun Y & Ayyildiz O. Two cases of bacterial meningitis accompanied by thalidome therapy in patients with multiple myeloma: is thalidomide associated with bacterial meningitis? Elsevier, 2008.

[8] Xiong Y. Artifical neuronal network based on principal component analysis input selection for clinical pattern recognition analysis, Elsevier.2007: 66-75

[9] Pingle M,Granger K, Feinberg P, ShatskyR, Sterling B, Rundell M, Spittzer E, larone D, Golightly L & Barany F. Multiplexed identification of blood-borne bacterial pathogens by use of a novel 16s rRNA gene PCR-ligase detection reaction capillary**.** electrophoresis assay. Microbiology, 1927-1935, 2007.

[10] Starosvetsky J,Starosvetsky E & and Armon R. Elecctrophoretic applications of sol-gel matrices. Ceramics international.2008: 1443-1448.

[11 Mao Y, Zhao Z, Wang S & Cheng Y. Urinary nucleosides based potential biomarker selection by support vector machine for bladder cancer recognition. analytical chemical, p. 43-40, 2007.

[12] G. Ceballos, J. Paredes, and L. Hernández. **A novel approach for pattern recognition in capillary electrophoresis dat**a. Bioengineering, 2007.

[13] Ton C, Yang C. Feature selection for SVM: An application to hypertension diagnosis. Expert system with application, p. 754-763, 2008.

[14] Lung C, Wang C. A GA-based feature selection and parameters optimization for support vector machine. Elsevier, p. 231-240, 2006.

[15] Máquez D, Paredes J. Nonlinear filters based on support vector machine. No. II, p. 581-584, 2007.

[16] Cala M, Vásquez A, Martínez J, Stasthenko E**.** Caracterización de compuestos fenólicos por electroforesis capilar de la especie phyllanthus acununatus (euphorbiaceae) y estudio de su actividad antioxidante. Scientia et Técnica.2007: 173-175.

[17] Martínez B, González L, Silva E, Montoya D, Morena R, Rada P, Hernández L**.** Concentración plasmática de serotonina y de arginina en pacientes con cefalalgias. Gac Med Caracas.2005:. 242-246.

[18] Solanilla L,Gómez C,Múnera L, Alineaminetos de mutiples secuencias**.** Sistemas y telemática**.** 2006.