Procesamiento de datos de espectrometría de masas: Algoritmos y metodologías

Sofía Calle, Silvia Chasiluisa, Jorge Carvajal, Roberto Herrera.
Facultad de Ingeniería Eléctrica y Electrónica, Escuela Politécnica Nacional, Quito, Ecuador.
Autores para correspondencia: jazmina.calle.jordan@gmail.com
Fecha de recepción: 28 de septiembre 2015 - Fecha de aceptación: 12 de octubre 2015.

RESUMEN
El cáncer es una enfermedad asintomática en una etapa temprana y muy difícil de diagnosticar. En muchos de los casos no es percibida hasta que ya alcanza la metástasis. Es la segunda causa de muerte en el Ecuador a pesar de que los avances conseguidos en los últimos tiempos han sido revolucionarios, existen casos en donde el cáncer es detectado en su etapa terminal y aún no se ha encontrado ninguna metodología científica ni empírica que indique la presencia de esta patología. Las metodologías tradicionales de diagnóstico en cualquiera de sus tipos aciertan a un número relativamente bajo de casos en sus etapas tempranas, usando métodos invasivos con el riesgo de ser falsos positivos o falsos negativos. Centros de investigación y universidades han juntado esfuerzos para buscar alternativas de diagnóstico y tratamiento del cáncer usando métodos que permitan mejorar la eficacia del diagnóstico. En este trabajo se aborda un análisis de las diferentes etapas implicadas en el procesamiento de datos de muestras de tejidos cancerosos y saludables usando espectrometría de masas, usando plataformas computacionales, aplicados al mejoramiento de la calidad de las mediciones para posteriores aplicaciones de definición de biomarcadores.
Palabras clave: Procesamiento de datos, espectrometría de masas, biomarcadores, plataformas computacionales, procesamiento digital de señales.

ABSTRACT
Cancer is an asymptomatic disease at an early stage and very difficult to diagnose in many cases it is not perceived until it has already reached the stage of metastasis, spreading in other organs of the body. It is the second cause of death in Ecuador despite progress made in recent years have been revolutionary, there are cases where the cancer is detected in its terminal stage and still has not found any scientific or empirical methodology to indicate the presence of this pathology. Traditional methods of diagnosing cancer in any of its types are relatively ineffective in early stages, using invasive methods and at risk of being detected as false positive or false negative. Research centers and universities have joined forces to seek alternative diagnosis and treatment of cancer using methods to improve the efficiency of diagnosis and detection of cancer in its early stages. This paper discusses a group of algorithms and methodologies for processing data sets of mass spectrometry measurements of cancer and normal analyzed samples using computing platforms aimed at improving the quality of measurements for biomarkers definition applications.
Keywords: Data processing, mass spectrometry, biomarkers, computing platforms, digital signal processing.

1. INTRODUCCIÓN

La Espectrometría de Masas (EM) es una técnica de adquisición de datos muy utilizada en investigaciones de enfermedades como el cáncer por ser capaz de extraer información y presenta una

gran facilidad de adaptación en plataformas computacionales, donde utilizando algoritmos de minería de datos y aprendizaje de máquina se están definiendo continuamente nuevas metodologías de diagnóstico del cáncer empleando biomarcadores (BM) como técnicas de detección temprana de cáncer.

La adquisición de datos se realiza a través de las técnicas de ionización como: *Electron Ionizations* (EI), *Fast Atom Bombardment* (FAB), *ElectroSpray Ionization* (ESI), *Matrix-assisted laser desorption/ionization* (MALDI) *y Surface-enhanced laser desorption/ionization* (SELDI) aplicada sobre una muestra de fluidos biológicos como saliva, orina o sangre y de esta forma se obtiene la información requerida (Kristjansdottir *et al*., 2013; Fishman, 1991; Cedazo-Minguez & Winblad, 2010). Esta técnica presenta sus mediciones en forma de señales discretas con una limitación al momento de procesar un gran volumen de datos representados en vectores y matrices con software computacional y análisis estadístico, razón por la cual se aplican algoritmos para eliminar datos redundantes y datos que carecen de información relevante. Además, los espectros obtenidos de la EM presentan varios problemas como heterogeneidad y tienen una mezcla de ruido de tipo eléctrico, químico, de procesamiento y ruido debido a la mala calibración de equipos por lo que es necesaria una etapa previa que elimine la mayor parte de este ruido. Las etapas a utilizar son: Remuestreo, Corrección de Línea de Base, Alineación, Normalización y Suavizado de Ruido (Kristjansdottir *et al.*, 2013; Ping, 2007).

Este documento se divide en varias secciones que ayuden al esclarecimiento de esta técnica de los cuales se mencionan así: la Sección 2 se habla de la importancia de la información de EM y como se definen los BM. Sección 3 la Adquisición de los Datos, conceptos básicos y necesarios como espectrómetro de masas y sus cuatro funciones internas principales. Sección 4 analiza las diferentes metodologías de procesamiento, objetivos, problemas y limitaciones. Sección 5 se mencionan las herramientas computaciones usadas para el procesamiento de datos. Al final se adjuntan las conclusiones de este trabajo.

2. INFORMACIÓN DE ESPECTROMETRÍA DE MASAS

La EM es una técnica analítica que permite medir de manera precisa el peso molecular de un compuesto, es un método muy versátil ya que permite identificar la estructura de varios tipos de compuestos. También se aplica a todo tipo de muestras de fluidos biológicos, volátiles, no volátiles, sólidos, líquidos o gaseosas. Los fluidos biológicos contienen proteínas que sirven para la identificación y búsqueda de BM. Los BM son cambios medibles provocados por sustancias ajenas al organismo, que indican el estado patológico o no patológico del ser humano. Para medir estos cambios se analizan los patrones de abundancias obtenidos de las mediciones de EM que definen proteínas y antígenos. Los antígenos son sustancias que produce el sistema inmunológico para la producción de anticuerpos contra virus, químicos o toxinas. Una aplicación importante del uso de BM es el diagnóstico, tratamiento y prevención de enfermedades de tipo cancerígeno en etapa temprana (Cedazo-Minguez & Winblad, 2010; van der Merwe *et al*., 2007; Martín Gómez & Ballesteros González, 2008; Diamandisi, 2004).

3. ADQUISICIÓN DE LOS DATOS

La EM produce información a partir de los iones generados de moléculas orgánicas en fase gaseosa. Estos iones producidos se separan de acuerdo a la relación masa/carga (*m/z*) y se contabiliza su intensidad (abundancia relativa) (Martín Gómez & Ballesteros González, 2008; Ping, 2007). A partir de los datos obtenidos se genera el espectro de masas, en el eje horizontal se representa la relación *m/z* [Th] (*Thomsons*) y en el eje vertical la abundancia relativa (Kristjansdottir *et al.*, 2013; Ping, 2007).

El instrumento denominado espectrómetro integra en su funcionamiento el proceso de adquisición de datos en formato digital. Dicho proceso se detalla en la Fig. 1, empieza con la

introducción de la muestra, ionización, analizador de masas, detección de iones. Los datos adquiridos son valores numéricos que al graficarlos toman la forma de una señal discreta.

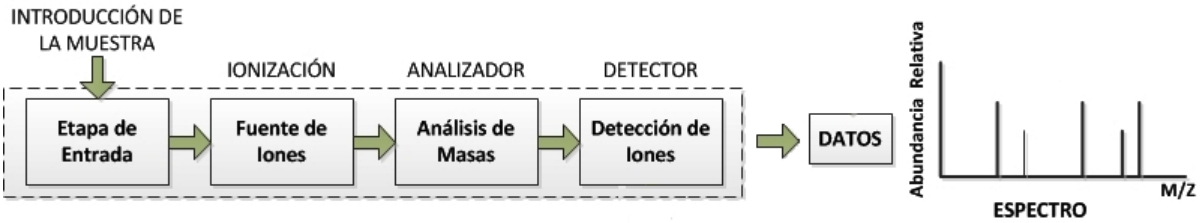

Figura 1. Esquema de un Espectrómetro de Masas**.**

3.1. Etapa de introducción de muestras

En esta etapa se toman pequeñas muestras como tejido, sangre o saliva para ser introducida en una cámara de volatilización en el vacío, donde con una fuente de calor se procede a cambiar el estado natural de la muestra (sólido o líquido) a estado gaseoso (Kristjansdottir *et al*., 2013; Fishman, 1991; Cedazo-Minguez & Winblad, 2010).

3.2. Etapa de ionización

La muestra en estado gaseosa es bombardeada con electrones, iones, moléculas o fotones, esto dependerá de la naturaleza de la muestra y el tipo de información que se desee obtener (Ping, 2007). Existen los siguientes métodos de ionización: Ionización en fase gaseosa e Ionización por desorción. En la ionización en fase gaseosa primero se volatiliza la muestra para luego ionizarla mientras que en la ionización por desorción la muestra se transforma directamente en iones. Se citan a continuación las técnicas por Desorción aplicadas en EM: Ionización por *electrospray* (ESI), Bombardeo con átomos rápidos (FAB) (van der Merwe *et al*., 2007; Martín Gómez & Ballesteros González, 2008; Gomis, 2008) y Ionización/Desorción por Láser (LDI).

3.3. Analizador de masas

Los iones atraviesan unos platos aceleradores que incrementan la energía cinética y pasan por un campo magnético que cambia la dirección de cada ion describiendo una curva que varía en función de su masa. Luego, dichos iones chocan sobre el detector de masas que contabiliza el número de colisiones en un punto específico. El número de colisiones se denomina abundancia relativa, de esta manera se obtiene el denominado espectro de masas con una resolución que varía en un rango de 10000 a 1000000 de puntos (van der Merwe *et al*., 2007; Gomis, 2008; Hilario *et al*., 2005). En la Fig. 2 se muestra un espectro de masas del conjunto de datos *Ovarian_Data_WCX2_CSV.zip*.

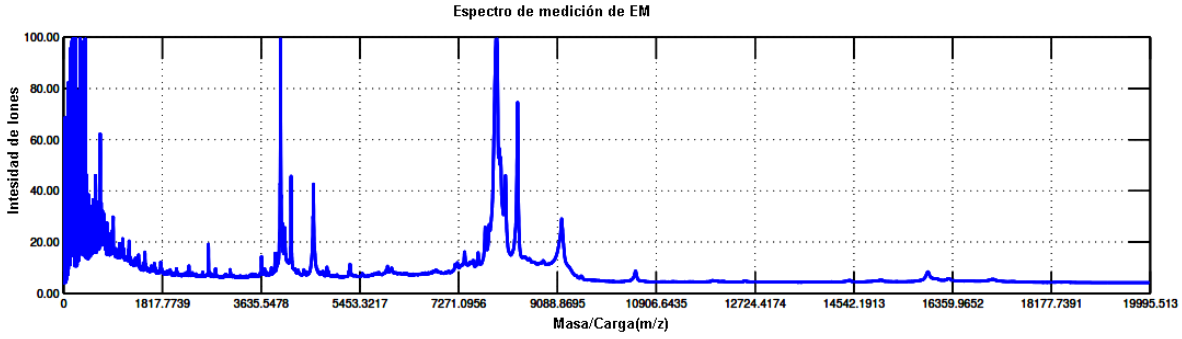

Figura 2. Ejemplo de un espectro de mediciones del Conjunto Ovarian Data WCX2 CSV.zip.

4. TIPOS DE METODOLOGÍAS

Durante el proceso de adquisición de los datos se introducen contaminantes debido a mala calibración de equipos, la preparación de la muestra, la inserción de la muestra en el instrumento y la saturación de iones. Estas variaciones y errores se traducen en ruido de diferentes características introducidos en los espectros, estos son: ruido térmico, eléctrico y químico. Además, los datos obtenidos tienen una heterogeneidad dimensional debido al tamaño de la muestra y la resolución de cada espectrómetro. Por ello una etapa previa al procesamiento de datos es extremadamente importante para extraer la señal de interés (Alterovitz & Ramoni, 2007). La metodología usada para procesar los datos son: Remuestreo, Corrección de Línea de Base, Alineación, Normalización y Suavizado de Ruido.

4.1. Remuestreo

Las aplicaciones prácticas con procesamiento digital de señales enfrentan el problema de cambiar la tasa de muestreo de la señal, ya sea aumentando o disminuyéndola, este paso se llama conversión de frecuencias de muestreo o remuestreo y en este caso se traduce como el aumento o disminución de la resolución de cada vector de *m/z*. El remuestreo es un proceso en el que se obtiene una nueva señal con valores controlados de *m/z*, esta nueva señal debe ser en lo posible similar a la original. Valores controlados significa que el número de puntos pueden ser mayores, iguales o menores al de la señal original. Esta etapa busca homogeneizar los vectores *m/z* para esto se aplica un factor *I/D*, donde I se denomina *Interpolation* que logra el incremento en la resolución de cada vector, y *D* se denomina *Decimation* este logra la disminución de la resolución de cada vector (Alterovitz & Ramoni, 2007). En la Fig. 3, ilustra el espectro de masas de un conjunto de datos Ovarian_Data_WCX2_CSV.zip remuestreado entre los valores de 2000 y 11000 (Ingle & Proakis, 2012).

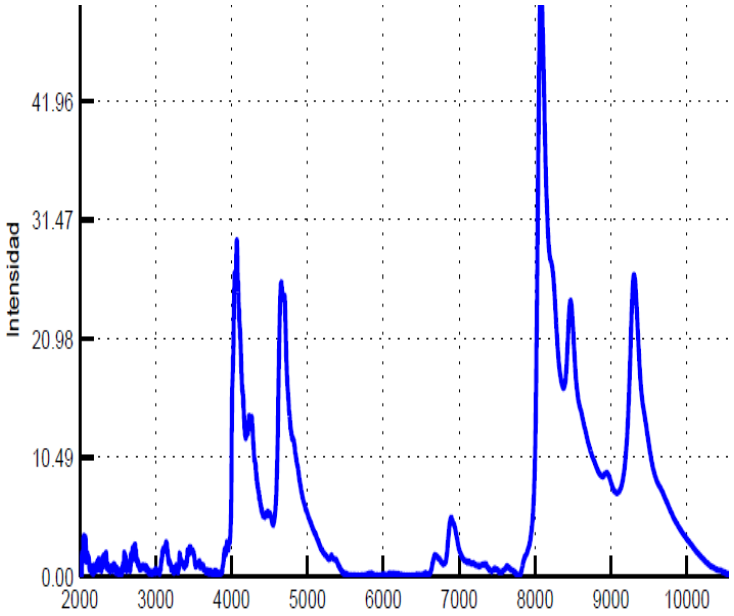

Figura 3. Remuestreo de los espectros.

4.2. Corrección de línea de base

Los datos de manera general muestran una línea de base variable consecuencia del ruido químico en la matriz o sobrecarga de iones que se origina en el detector de iones cuando este se satura. La línea de base es un desplazamiento de los iones en el eje vertical que eleva los valores de *m/z* bajos mientras que los valores altos de *m/z* no se ven tan afectados tal como se muestra en la Fig. 4. Para corregir este problema se estima la línea de base y se resta del espectro original es decir se halla el punto más bajo del espectro y se arrastra hasta cero en el eje vertical (Alterovitz & Ramoni, 2007; Antoniadis *et al.*, 2010).

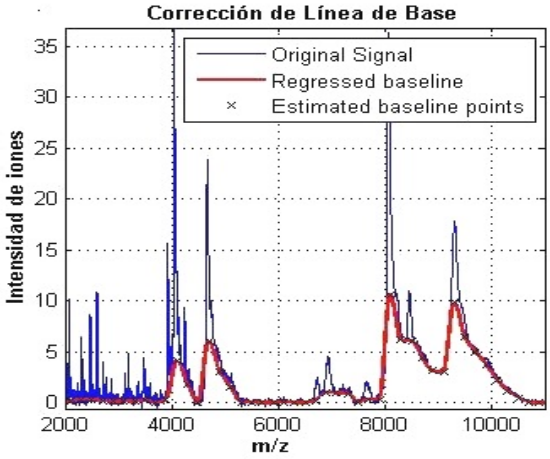

Figura 4. Estimación Línea de Base.

Se han desarrollado varias técnicas que ayudan a estimar la línea de base como: Filtros pasa altos implementados con transformada rápida de Fourier, teoría de *wavelets* y filtros digitales. Los métodos mencionados son poco usuales ya que estos distorsionan la señal y sería necesario modelar un filtro para cada espectro. El algoritmo más utilizado para estimar la línea de base en EM se basa en interpolación *spline* o suavizado (Alterovitz & Ramoni, 2007; Hilario *et al*., 2005, Eidhammer & Mikalsen, 2007).

Spline es una función polinómica a trozos de grado *p*, siendo la más práctica el polinomio de grado 3. Se definen el número y posición de los nodos, los nodos dividen al espectro en regiones en el eje de *m/z* (intervalos). El modelo matemático para un polinomio cúbico *Spline* se muestra en la ecuación 1.

$$s(x) = A_i(x - x_i)^3 + B_i(x - x_i)^2 + C_i(x - x_i) + D_i \quad para \quad i = 0,...,n-1 \tag{1}$$

Donde, A_i, B_i, C_i, D_i, $\epsilon\ \mathbb{R}$, $x\ \epsilon(x,x_i)$ son las frontera de cada intervalo y n es el grado del polinomio. Se trata de estimar A_i, B_i, C_i, D_i, principalmente con condiciones de continuidad (s, s', s") y condiciones de interpolación en los nodos. Se repite el paso anterior para cada intervalo de la señal. Y se construye la línea de base. Se resta la línea de base estimada del espectro original (Alterovitz & Ramoni, 2007; Gustafsson *et al*., 2011; Capelo-Martínez *et al*., 2015).

4.3. Alineación

La etapa de alineación de picos es de suma importancia debido a que existe una variación significativa entre las muestras de la intensidad y la ubicación de los picos en *m/z* por la mala calibración de los espectrómetros de masas. La idea es reemplazar los valores originales de *m/z* por valores calibrados o alineados, definiendo un vector de picos con los valores máximos de intensidades. Los picos no alineados se desplazan hacia las zonas donde hay más alineación, obteniendo así nuevos vectores de intensidades (Alterovitz & Ramoni, 2007; Bachmayer, 2007). En la Fig. 5 en a) Se muestra el mapa de calor de los datos con sus picos distorsionados sobre el eje *x* y en la parte b) Se observa el espectro con los picos alineados, estos de concentran en una línea vertical para valores de *m/z* de 4000, 8000 y 9000 aproximadamente.

4.4. Normalización

La normalización se realiza para que los diferentes espectros sean comparables entre sus intensidades relativas. Este método se utiliza para identificar y eliminar las variaciones aleatorias en la amplitud de cada intensidad causadas por la mala calibración de los instrumentos. En esta etapa busca reducir las diferencias de las intensidades para cambiar la escala. Se localiza el valor máximo de intensidad para

asignarle un valor y el resto de valores se ajusten proporcionalmente a este (Alterovitz & Ramoni, 2007). En la siguiente ecuación 2, se muestra el factor de intensidad normalizada.

$$I_{N\,m*n} = \frac{I_{i*n}}{I_{MAX\,i*n}} * N_f \qquad (2)$$

Donde $I_{N\,m*n}$ es la matriz normalizada, I_{i*n} es cada intensidad a re-escalar, $I_{MAX\,i*n}$ es la intensidad máxima en cada espectro y N_f es el factor de normalización (Eidhammer & Mikalsen, 2007).

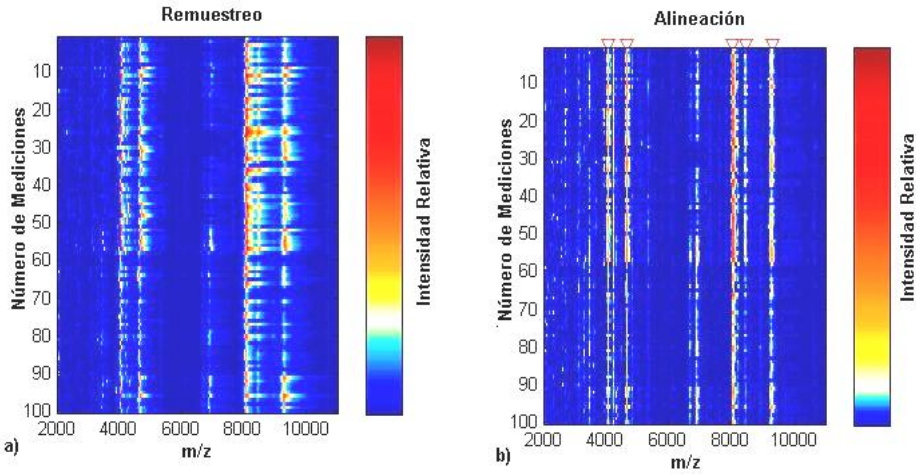

Figura 5. Alineación de picos del Conjunto Ovarian_Data_WCX2_CSV.zip.

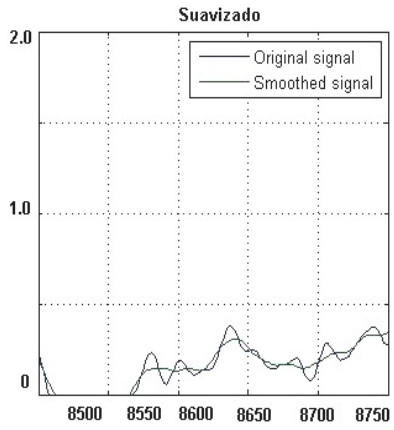

Figura 6. Suavizado de un espectro.

4.5. Suavizado de ruido

En esta etapa se busca reducir el ruido producido en etapas anteriores. Para ello se aplica técnicas de suavizado que produce una curva más suave del espectro reduciendo al máximo los picos falsos. Esta proceso se lleva a cabo utilizando el Filtro de *Savitzky y Golay*, consiste en suavizar muestra a muestra la señal basándose en una regresión polinomial (Alterovitz & Ramoni, 2007). Este tipo de filtro se adapta a la variación de frecuencia de muestreo y conserva la agudeza de los picos. El método de suavizado polinómico de mayor aceptación es el de *Savitzky-Golay* y *Káiser* que emplea un filtro digital de polinomio de mínimos cuadrados el cual conserva la mayor parte de las características de la señal así como la resolución entre picos de iones y la altura de los picos. Sin embargo este tipo de algoritmos requiere un análisis más exhaustivo de software (Alterovitz & Ramoni, 2007; Bachmayer,

2007). En la Fig. 6, se muestra el suavizamiento de picos, eliminando cambios bruscos entre cada pico para mayor apreciación se amplía la zona de interés para valores de *m/z* de 8500 a 8750. En la Fig. 7 se muestra el espectro de masas de un conjunto de datos *Ovarian_Data_WCX2_CSV.zip*, producto del procesamiento. En lo posible se ha reducido el ruido.

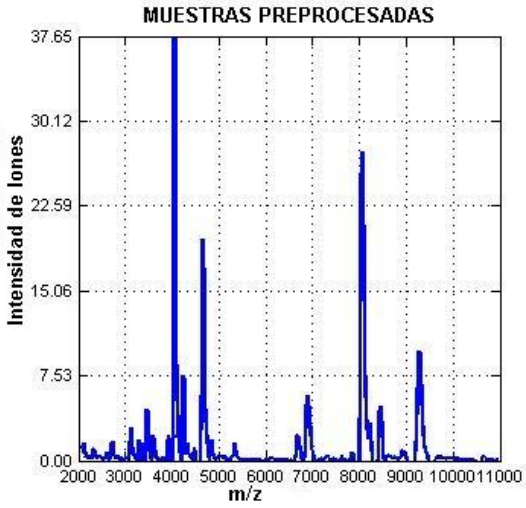

Figura 7. Espectro de masas después del procesamiento.

5. HERRAMIENTAS COMPUTACIONALES

5.1. Matlab bioinformatics toolbox

Esta poderosa herramienta ofrece varios algoritmos para el análisis de microarrayas, espectrometría de masas y la oncología de los genes. En el campo de la EM incluye el procesamiento con corrección de línea de base, suavizado, calibración y remuestreo. Adicional, proporciona funciones para la clasificación e identificación de biomarcadores potenciales en datos adquiridos a través de SELDI y MALDI. Además ofrece funciones como *heatmap* espaciales, navegadores de secuencias y *clustergrams* para visualizar los datos (Guide, 2003).

5.2. Weka3

Este es un software exclusivo de minería de datos desarrollado en JAVA. Integra en su plataforma un conjunto de algoritmos de aprendizaje automático como pre-procesamiento, clasificación de características, regresiones, \textit{clustering} y reglas de asociación. Las principales características de Weka son:

- Es un software libre publicado bajo licencia GNU con plataforma amigable para personas que no conocen a fondo la minería de datos (Cannataro *et al*., 2005; Markov & Russell, 2011).
- Para el pre-procesamiento de los datos es necesario definir el origen de los datos, weka 3 admite los siguientes formatos: .arff por defecto, csv (archivos separados por comas o tabuladores), c4.5 (conformado por el fichero .names y el fichero .data) (Cannataro *et al*., 2005; Markov & Russell, 2011).

5.3. Mass-Up

Es un software de *Open Source* para el análisis de datos obtenidos con la técnica de EM por MALDI desarrollada en java. Además de ser un software libre es capaz de cargar datos de MALDI desde diferentes formatos como mzML, mzXML y CSV. Esta herramienta posee 4 secciones para mejorar la

interacción con el usuario las cuales son: menú de carga, menú de preproceso, menú de análisis y principales tipos de datos.

En la sección del pre procesamiento destacan las opciones de método de suavizado por el método de *Savitzky Golay*, método de corrección de línea de base, detección de picos e intensidad mínima de pico. Adicional, en el sitio web de descarga del software se encuentran varios archivos de datos con los que se puede realizar varias pruebas para probar su funcionamiento (Capelo-Martínez *et al*., 2015).

6. CONCLUSIONES

- Las mediciones de espectrometría de masas poseen características que requieren algoritmos y metodologías de procesamiento específicas para mejorar los problemas de ruido, contaminantes, efecto de línea de base y diferencias dimensionales de las mediciones.
- El procesamiento digital de señales de mediciones de espectrometría de masas engloba una sinergia de técnicas estadísticas, algoritmos computacionales y conceptos de procesamiento digital de señales, lo que lo hace específico y dedicado para estas aplicaciones. Las características de estas mediciones difieren sustancialmente de las señales tradicionalmente analizadas en teoría de comunicaciones.

REFERENCIAS

Alterovitz, G., M.F. Ramoni, 2007. *Systems bioinformatics: An engineering case-based approach.* Boston-London: TRECH House, Inc.

Antoniadis, A., J. Bigot, S. Lambert-Lacroix, 2010. *Peaks detection and alignment for mass spectrometry data*. Disponible en http://membres-timc.imag.fr/Sophie.Lambert/papier/Spectrometry.pdf, 19 pp.

Bachmayer, S., 2007. *Preprocessing of mass spectrometry data in the field of proteomics*. Disiponible en http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.486.6689&rep=rep1&type=pdf, 17 pp.

Cannataro, M., P.H. Guzzi, T. Mazza, P. Veltri, 2005. *Preprocessing, management, and analysis of mass spectrometry proteomics data*. Università Magna Græcia di Catanzaro, Italy. Disponible en http://www.researchgate.net/profile/Tommaso_Mazza/publication/255586127_Preprocessing_Management_and_Analysis_of_Mass_Spectrometry_Proteomics_Data/links/0a85e5350cae8411ec000000.pdf, 5 pp.

Capelo-Martínez, J.L., F. Fdez-Riverola, D. Glez-Peña, A. Gutiérrez Jácome, H. López-Fernández, E. Lorenzo Iglesias, J.R. Méndez Reboredo, R. Pavón Rial, M. Reboiro-Jato, 2015. *Manual de mass up tEAM*. Disponible en http://sing.ei.uvigo.es/mass-up/manual.

Cedazo-Minguez, A., B. Winblad, 2010. Biomarkers for Alzheimer's disease and other forms of dementia. *Exp Gerontol*., 45(1), 5-14.

Cheng, Y., 2009. *Analysis of Seldi mass spectra for biomarker discovery and cancer classification*. PhD dissertation, CRUK Cancer Studies, Medical School, University of Birmingham. Disponible en http://etheses.bham.ac.uk/317/1/Cheng09Phd.pdf, 244 pp.

Diamandisi, E.P., 2004. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool. *Mol. Cell Proteomics*, 3(4), 367-78.

Eidhammer, I., K. Flikka, L. Martens, S-O. Mikalsen, 2007. Computational methods for mass spectrometry proteomics. Wiley Online Library. Disponible en http://onlinelibrary.wiley.com/book/10.1002/9780470724309, 284 pp.

Fishman, D.A., 1991. *National ovarian cancer early detection program*. Mount Sinai School of Medicine. Disponible en https://www.mountsinai.org/static_files/MSMC/Files/Patient%20Care/OBGYN%20and%20Reproductive%20Services/NOCEDP%20Brochure.pdf, 20 pp.

Gomis, V.Y., 2008. *Espectrometría de masas*. Universidad de Alicante. Departamento de Ingeniería Química. Disponible en http://hdl.handle.net/10045/8249.

Guide, B.T.U., 2003. *Statistics Toolbox User's Guide*. The MathWorks Inc., Natick, MA, 816 pp.

Gustafsson, J.O.R., M.K. Oehler, A. Ruszkiewicz, S.R. McColl, P. Hoffmann, 2011. MALDI Imaging Mass Spectrometry (MALDI IMS) - Application of Spatial Proteomics for Ovarian Cancer Classification and Diagnosis. *Int. J. Mol. Sci*., 12(1), 773-794.

Hilario, M., A. Kalousis, C. Pellegrini, M. Müller, 2005.Processing and classification of protein mass spectra. *Mass Spectrom Rev.*, 25(3), 409-49.

Ingle, V. K., J.G. Proakis, 2012. Digital signal processing using MATLAB. CENGAGE Learning. BookWare Companion Series. Disponible en http://www.ece.iit.edu/~biitcomm/Yarmouk/Digital%20Signal%20Processing%20Using%20Matlab%20v4.0%20(John%20G%20Proakis).pdf, 816 pp.

Kristjansdottir, B., K. Levan, K. Partheen, E. Carlsohn, K. Sundfeldt, 2013. Potential tumor biomarkers identified in ovarian cyst fluid by quantitative proteomic analysis, iTRAQ. *Clin. Proteomics*, 10(1), 4.

Markov, Z., I. Russell, 2011. *An introduction to the WEKA data mining system*. Central Connecticut State University. Disponible en http://www.cs.ccsu.edu/~markov/weka-tutorial.pdf, 60 pp.

Martín Gómez, C., M. Ballesteros González, 2008. *Espectrometría de masas y análisis de biomarcadores*. Monografías de la Real Academia Nacional de Farmacía. Disponible en http://www.analesranf.com/index.php/mono/article/view/1066, 56 pp.

Ping, H., 2007. *Classification methods and applications to mass spectral data*. Hong Kong Baptist University. Restricted Access Theses and Dissertations. Paper 593.

van der Merwe, D.E., K. Oikonomopoulou, J. Marschall, E.P. Diamandis, 2007. Mass spectrometry: uncovering the cancer proteome for diagnostics. *Adv. Cancer Res*., 96, 23-50.