Identificación de longitudes de onda en las regiones NIR y MIR para la medición no invasiva de glucosa en sangre

Wavelength identification in NIR and MIR regions for non invasive blood glucose measurement

I. D. Castro, J. E. Vargas, F. Fonthal^(*)

Advanced Materials for Micro and Nanotechnology –Grupo IMAMNT–, Facultad de Ingeniería, Universidad Autónoma de Occidente, Calle 25 N° 115-85, Cali, Colombia.

^(*) Email: ffonthal@uao.edu.co

Recibido / Received: 15/05/2012. Revisado / Revised: 31/07/2012. Aceptado / Accepted: 28/08/2012.

DOI: <http://dx.doi.org/10.7149/OPA.45.3.323>

RESUMEN:

En este documento se presenta un análisis cuantitativo de espectros en las regiones del Infrarrojo cercano (NIR) y medio (MIR) correspondientes a muestras de diferente complejidad (agua, plasma y sangre entera) con concentraciones conocidas de glucosa, con el fin de identificar las características de procesamiento y regresión óptimas para la medición de glicemia. Utilizando técnicas de filtrado y corrección de línea base para los espectros, algoritmos de extracción de parámetros (Información Mutua y Factor de Mérito) y de regresión para construcción de modelos matemáticos (PLS, LR, PCR), se obtuvieron 180 modelos que fueron evaluados a partir del error de validación cruzada (RMSECV). Se encontraron diferentes longitudes de interés en las regiones comprendidas entre 648 cm⁻¹ y 1250 cm⁻¹, y entre 2350 cm⁻¹ y 2565 cm⁻¹ (región MIR), y entre 11100 cm⁻¹ y 11950 cm⁻¹, y entre 5190 cm⁻¹ y 5700 cm⁻¹ (Región NIR).

Palabras clave: Glucosa en sangre, Espectroscopía IR, Transmitancia, Medición no Invasiva, Información Mutua, Factor de Mérito.

ABSTRACT:

This document presents a quantitative analysis of Near Infrared (NIR) Region and Medium Infrared (MIR) Region spectra corresponding to increasing complexity matrices (water, plasma and whole blood) with known concentrations, in order to identify processing and regression characteristics to obtain a good blood glucose measurement. Using filtering and baseline correction techniques, parameter extraction algorithms (Mutual Information and Merit Factor), and regression algorithms (PLS, LR, PCR), 180 models were obtained and tested using root mean square error of cross validation (RMSECV). In the MIR spectrum, wavelengths between 648 cm⁻¹-1250 cm⁻¹ and between 2350 cm⁻¹-2565 cm⁻¹ were identified as potential predictors, while in the NIR spectrum wavelengths between 11100 cm⁻¹-11950 cm⁻¹, and between 5190 cm⁻¹-5700 cm⁻¹ were outstanding.

Key words: Blood Glucose, IR Spectroscopy, Transmittance, Non-Invasive Measurement, Mutual Information, Merit Factor.

REFERENCIAS Y ENLACES / REFERENCES AND LINKS

- [1]. J. G. Webster, E. R. Ritenour, S. Tabakov, N. G. Kwan-Hoong, *Handbook of Optical Sensing of Glucose in Biological Fluids and Tissues*, CRC Press, Boca Raton, pp. 1-40 (2009).
- [2]. M. Duarte, G. Muñoz, J. Rodríguez, A. Escorza, "Prevalencia, detección y tratamiento de la diabetes gestacional", *Revista Salud Pública y Nutrición* 5, No. 1 (2004). Disponible en: http://www.respyn.uanl.mx/v/1/ensayos/diebetes_gest.htm Revisado en julio de 2011.

- [3]. A. González, J. L. Rosenzweig, G. Umpierrez, "Self-monitoring of blood glucose", *J. Clinical Endocrinology and Metabolism* **92**, No. 5 p.0 (2007).
- [4]. P. Restrepo, "Glucómetro no invasivo", *Revista de la Facultad de Medicina* **10**, 22-27 (2005).
- [5]. L. Pulenta, E. Ávila, "Medición no invasiva de glucosa por infrarrojo cercano". XV Congreso Argentino de Bioingeniería, Argentina 047ei (2005).
- [6]. C. Araujo-Andrade, F. Ruiz, J. R. Martínez, H. Terrones, "Predicción no invasiva de los niveles fisiológicos de glucosa utilizando espectroscopia de absorción infrarroja (NIR)", Congreso Anual de la AMCA, México, pp. 281-285 (2004).
- [7]. J. E. Vargas, D. Montaña, Estudio sobre la Utilización de Espectroscopía Infrarrojo para Medir la Concentración de Glucosa en Sangre, Trabajo de Grado, Ingeniero Biomédico, Universidad Autónoma de Occidente, Facultad de Ingeniería, Santiago de Cali, Colombia (2009).
- [8]. J. G. Webster, E. R. Ritenour, S. Tabakov, N. G. Kwan-Hoong, Handbook of Optical Sensing of Glucose in Fluids and Tissues, CRC Press, Boca Raton, pp. 97-114 (2009).
- [9]. J. E. Vargas, I. D. Castro, F. Fonthal, "Identificación espectroscópica en la región media y cercana del infrarrojo para su aplicación en mediciones no invasivas de glucosa", V Congreso Latinoamericano de Ingeniería Biomédica, CLAIB 2011, La Habana, Cuba. IFMBE Proceedings, 33, Id Trabajo: 226 (2011).
- [10]. I. Gabriely, "Transcutaneous glucose measurement using near-infrared spectroscopy during hypoglycemia", *Diabetes Care* **22**, 2026-2032 (1999).
- [11]. Y. C. Shen, A. G. Davies, E. H. Linfield, "Determination of glucose concentration in whole blood using Fourier-Transform Infrared spectroscopy", *J. Biol. Phys.* **29**, 129-133 (2003).
- [12]. A. Savitzky, M. Golay, "Smoothing and differentiation of data by simplified least squares procedures", *Anal. Chem.* **36**, 1627-1639 (1964).
- [13]. http://www.mathworks.com/help/techdoc/data_analysis/bqm3i7n-13.html
- [14]. T. Lan, Y. Fang, W. Xiong, C. Kong, "Automatic baseline correction of infrared spectra", *Chin. Opt. Lett.* **5**, 613-616 (2007).
- [15]. A. Candolfia, R. De Maesschalck, D. Jouan-Rimbaud, P. A. Haineyb, D. L. Massart, "The influence of data pre-processing in the pattern recognition of excipients near-infrared spectra", *J. Pharmaceut. Biomed.* **21**, 115-132 (1999).
- [16]. M. A. Romero, Desarrollo de Nuevas Metodologías Analíticas en el Control de Calidad de la Industria Farmacéutica, PhD Thesis in Chemistry, Universidad Autónoma de Barcelona (Spain), Department of Chemistry (2001).
- [17]. F. M. González, Desarrollo de un Método de Selección de Variables para Datos Espectroscópicos en el Infrarrojo Cercano, Trabajo de Grado, Ingeniero Electrónico, Universitat Rovira i Virgili de Tarragona (Spain), Departamento de Ingeniería Electrónica, Eléctrica y Automática (2004).
- [18]. N. Benoudjit, D. Francois, M. Meurens, M. Verleysen, "Spectrophotometric variable selection by mutual information", *Chemometr. Intel. Lab.* **74**, 243-251 (2004).
- [19]. F. Rossi, A. Lendasse, D. Francois, V. Wertz, M. Verleysen, "Mutual information for the selection of relevant variables in spectrometric nonlinear modelling", *Chemometr. Intel. Lab.* **76**, 215-226 (2006).
- [20]. Y. J. Kim, G. Yoon, "Prediction of glucose in whole blood by near-infrared spectroscopy: Influence of wavelength region, preprocessing, and haemoglobin concentration", *J. Biomed. Opt.* **11**, 041128 (2006).
- [21]. C. Petibois, "Determination of glucose in dried serum samples by Fourier transform infrared spectroscopy", *Clin. Chem.* **45**, 1530-1535 (1999).
- [22]. Y. J. Kim, S. Hahn, G. Yoon, "Determination of glucose in whole blood samples by mid-infrared spectroscopy", *Appl. Opt.* **42**, 745-749 (2003).

1. Introducción

Las personas con alteraciones en los niveles de glucosa necesitan de una medición frecuente de los mismos, con el fin de conocer su estado actual y realizar un control adecuado para

prevenir las complicaciones derivadas de patologías como la diabetes mellitus (DM) que pueden poner en riesgo la vida [1].

Las personas que padecen DM, se ven obligadas a realizar chequeos de sus niveles de

glucosa en sangre con una periodicidad que oscila entre 3 y 5 veces al día [2,3]. Esto mediante la utilización de glucómetros convencionales que utilizan una muestra de sangre extraída por punción de la yema de los dedos; lo cual es una práctica dolorosa e incómoda que daña los tejidos, limita un control adecuado de la glicemia y las complicaciones de la DM, y aumenta los riesgos de infección en estas personas [1].

La utilización de técnicas invasivas junto con la necesidad de la obtención de estas mediciones en personas con alteraciones en glicemia, han llevado a realizar estudios acerca de diferentes técnicas para mediciones no invasivas, indoloras, cómodas y prácticas para el usuario [4,5]. Algunas incluyen la iontoforesis reversa, el monitoreo continuo mediante biosensores implantados, fotoacústica, impedanciometría de ondas de radio, rotación de luz polarizada, emisión termal y espectrofotometría; cada una con limitantes que deben ser solucionados para su aplicación clínica [4,5].

En la actualidad, la técnica más aceptada científicamente es la espectrofotometría [6], la cual estudia la absorción de la luz por parte de las moléculas presentes en una muestra, y mediante este fenómeno busca la determinación cuantitativa de los compuestos en la solución analizada, tal y como lo realiza la medición no invasiva de la saturación de Oxígeno en sangre (SP02) utilizando una pinza en el dedo índice

Debido a esto, esta técnica ha sido ampliamente estudiada con el fin de determinar las regiones del espectro y las posibles longitudes de onda que interactúan de una mejor forma con la molécula de glucosa, coincidiendo en la utilización de luz infrarroja para este fin [6-8].

Es de aclarar que el parámetro que se estudia mediante esta técnica es la cantidad de la molécula D-glucosa $C_6H_{12}O_6$ disuelta, teniendo como referencia la cuantificación por métodos estándar (Glucómetro convencional). De acuerdo a esto, este tipo de estudios no incluyen un análisis de los compuestos que son resultado del proceso de glicosilación como la hemoglobina glicada HbA1c, pues el parámetro bajo el cual se

pretende desarrollar una medición es la glicemia.

A partir de estos hallazgos, se han realizado investigaciones con el fin de mejorar, desde una perspectiva instrumental y de procesamiento de datos, la predicción de los niveles de glucosa en sangre. Estas últimas han concluido en la necesidad de mejorar este tipo de métodos para su aplicación clínica, mediante estudios enfocados hacia las técnicas de procesamiento de las señales y la selección de la región de análisis [6,7,9].

A pesar de esto, existe una discrepancia en las regiones del espectro infrarrojo más adecuadas para este fin. Se identifican estudios que utilizan grupos de longitudes de onda contiguas en la región cercana del espectro infrarrojo (NIR por sus siglas en inglés: *Near Infrared*) [6,10] o diversas longitudes de onda aisladas en la misma región [5]; mientras otros han seleccionado longitudes en la región media del espectro infrarrojo (MIR por sus siglas en inglés: *Medium Infrared*) [7,11]. Estos estudios utilizan diversas técnicas de procesamiento con el fin de mejorar los errores de predicción.

Estas diferencias técnicas y metodológicas entre los resultados de los diferentes investigadores, junto con la dificultad que implica que los picos de absorción de diferentes compuestos biológicos como la sangre y el tejido se solapan entre sí e intervienen en la medición, no han permitido el desarrollo de un dispositivo completamente no invasivo mediante técnicas de espectrofotometría en la región infrarroja. Esto hace necesario retomar el proceso y llevar a cabo estudios detallados in-vitro que incluyan diferentes técnicas de procesamiento de señales, para que una vez establecidas las características del sistema de medida se realicen pruebas similares utilizando potenciales interferentes biológicos, y pruebas in-vivo en sujetos sanos y con DM.

Debido a lo anterior, este trabajo tuvo su enfoque en el análisis in-vitro de D-glucosa disuelta en diversas matrices a través de espectros en las regiones NIR y MIR de forma simultánea. Esto con el fin de identificar las longitudes de onda útiles en ambos espectros para la determinación de la concentración de

glucosa y realizar una comparación con los estudios anteriores.

Se prepararon muestras de creciente complejidad (matrices de agua, plasma y sangre), con el objetivo de observar el efecto de la adición de interferencias biológicas y evidenciar un posible solapamiento de picos, al evaluar el desempeño de los modelos de predicción. De esta forma se identificaron las regiones de los espectros analizados que arrojaron una mayor correlación con la concentración de glucosa, evaluándolas mediante el error estándar medio de validación cruzada (RMSECV) de los modelos.

La selección o identificación de las regiones del espectro infrarrojo se realizó mediante técnicas de selección que ponderan cada longitud de onda de acuerdo a su relación con la concentración del analito utilizando dos métricas diferentes: el factor de mérito, y la información mutua. Estas técnicas brindan información acerca de la potencial utilidad de cada una de las longitudes de onda en la predicción de la molécula, por lo que permitieron realizar la selección de las longitudes de onda a utilizar en los modelos de regresión de tres formas diferentes..

2. Metodología

2.a. Diseño experimental

Se prepararon muestras con diferentes concentraciones que permitieran llevar a cabo el estudio en el rango fisiológico de la glucosa en sangre (20 mg/dl – 260 mg/dl) [1]. Las matrices utilizadas para las muestras de concentraciones crecientes fueron agua, plasma y sangre, para las cuales se utilizaron técnicas de dilución y adición estándar.

Estos 3 tipos de muestras fueron incluidos en el estudio debido a la necesidad de identificar las longitudes de onda sin ningún interferente biológico, y después evaluar los resultados de incluir los interferentes presentes en el plasma y la sangre.

Empleando la técnica de dilución se construyeron 17 muestras de agua con D-glucosa con concentraciones diferentes en el

rango especificado, en intervalos de 10 mg/dl y 20 mg/dl. Estas muestras fueron preparadas por triplicado a partir de una solución base de 400 mg/dl con el objetivo de disminuir los errores involucrados en la construcción de las muestras. Este procedimiento se realizó dos veces en diferentes días con lo que se elaboraron en total 102 muestras de D-glucosa en agua para la obtención de los espectros correspondientes en las regiones NIR y MIR.

Por su parte, para las muestras de plasma (80% de plasma) y sangre (71,42% de sangre), se realizó un procedimiento de adición estándar en el que se utilizaron muestras de dos sujetos sanos en ayunas, previa aceptación por consentimiento informado. Se mezclaron estas muestras con volúmenes específicos de soluciones de D-glucosa y agua y con agua ultra pura, para un incremento teórico de 20 mg/dl a partir de la concentración de la muestra con adición única de agua, obteniendo como resultado 11 muestras de plasma y 11 de sangre con adiciones desde 0 mg/dl (correspondiente a dilución de la muestra) hasta 200 mg/dl.

Para la extracción de las muestras de plasma y sangre, se utilizó el sistema de venoclisis por vacío BD Vacutainer y se almacenaron las muestras iniciales en tubos de ensayo con Heparina de Sodio para las muestras destinadas a la extracción de plasma y con Oxalato de Potasio y Fluoruro de Sodio para las muestras de sangre. Tanto las mediciones iniciales de las concentraciones de glucosa de los sujetos en ayunas como los resultados del proceso de adición estándar se midieron tres veces utilizando el glucómetro Contour TS de Bayer, para obtener la concentración final real a la que se llegó con cada una de las adiciones.

Los espectros de transmitancia se obtuvieron utilizando el espectrofotómetro Shimadzu Prestige IR-21, y los accesorios ATRMax II (Pike Technology) para la región MIR, y Demountable Liquid Cell (Pike Technology) para la región NIR. Mediante orden aleatorio, indicado por el software Matlab®, se tomaron los espectros de las muestras

Se tomaron tres espectros para cada una de las muestras, en momentos diferentes según lo indicó el proceso de aleatorización, con el fin de

evitar la influencia de condiciones temporales en el ambiente del laboratorio.

Es así que se prepararon 6 sets de muestras (2 de Agua, 2 de Plasma y 2 de Sangre), para ser analizados en las regiones NIR y MIR mediante la obtención de tres espectros por cada una de las muestras. Obteniendo un total de 12 sets de espectros en los cuales cada muestra tiene tres espectros

2.b. Procesamiento y resumen de datos

Cada una de las señales de transmitancia fueron filtradas mediante el método Savitzky-Golay utilizando el software MatLab®. Esto con el fin de tener una señal con la menor distorsión posible para la posterior extracción de información analítica. Éste método es comúnmente utilizado en señales de espectroscopía, debido a que se conserva la información de alta frecuencia característica de estas señales [12].

La técnica de promediado de espectros es ampliamente utilizada en el área de quimiometría que busca reducir el ruido presente en las señales obtenidas e identificar los denominados “outliers” dentro del set de medición. Con el fin de tomar ventaja de la repetición (tres veces de forma aleatoria) de los espectros de transmitancia y de la preparación por triplicado de las muestras de agua y glucosa, se realizó una identificación gráfica de outliers que eliminó espectros alterados por posibles factores aleatorios (sin tener que eliminar ninguna de las concentraciones por completo), seguida de un promediado de los espectros restantes correspondientes a cada una de las concentraciones.

Durante este proceso se identificó que el set de espectros obtenidos en la región NIR para uno de los dos grupos de muestras de agua y glucosa evidenció un comportamiento completamente diferente al de los otros espectros tomados con la misma técnica, por lo que fueron eliminados del estudio.

De igual forma se descartaron espectros individuales de algunas muestras sin eliminar ninguna de las concentraciones, pues los espectros restantes fueron promediados para obtener el espectro final de cada una. Una vez

realizado el promediado, se obtuvo un total de 139 espectros para ambas regiones, a los cuales se les aplicaron las técnicas de procesamiento espectral planteadas.

Buscando identificar las técnicas de procesamiento que llevaran a una mejor predicción de la glucosa en sangre, se aplicaron de forma paralela cinco procesamientos en los que se incluyeron correcciones de línea base: Sustracción de modelado lineal (Detrend) [13]; Sustracción de modelado iterativo polinomial con recorte [14]; Primera y segunda derivada de la señal; y técnica de variable normal estándar (SNV) [15,16].

De esta forma, se obtuvieron 6 sets de señales para cada una de las diferentes matrices (agua, plasma sangre), incluyendo los resultados de los procesamientos mencionados anteriormente y las señales originales filtradas. En la Fig. 1 es posible evidenciar el esquema de este procesamiento inicial y los 6 sets de datos como resultado.

2.c. Selección de regiones y longitudes de onda

Con el fin de realizar una identificación de longitudes de onda y regiones que pudieran llevar a una predicción acertada de la concentración de glucosa de las muestras, se aplicaron tres métodos de selección de variables basados en técnicas estadísticas utilizadas para determinar la importancia de una variable determinada en relación a la predicción de una segunda, pero que no habían sido aplicadas aún por los estudios orientados hacia la identificación de longitudes de onda IR para la predicción de glucosa en sangre.

La primera de las técnicas es el cálculo del factor de mérito (FM) [17]. Este cálculo se lleva a cabo a partir de la varianza de los espectros obtenidos para cada longitud de onda; El método realiza una comparación entre varianza interclase e intraclase, para lo que se agrupan espectros definiendo rangos de concentraciones de la muestra a analizar.

Un valor alto de FM a una longitud de onda particular implica que la distancia media entre cada medida de una clase y la posición del

Fig. 1: Esquema de procesamiento de las señales de transmitancia en las regiones NIR y MIR.

centroide de la misma es menor a la distancia media entre las centroides de todas las clases. Éste índice determina si la variabilidad observada corresponde al azar, o si existen diferencias significativas de la señal para cada una de las muestras y permite realizar una selección de las variables con mayor *FM*. El cálculo se realiza a partir de la Ec. (1):

$$FM_j = \frac{V.interclase}{V.intraclase} \quad (1)$$

Por su parte, la segunda técnica implica el cálculo de la información mutua (*IM*) [18,19]. Ésta métrica, al igual que la anterior, permite identificar la importancia de cada variable para la predicción a realizar; sin embargo, se basa en técnicas estadísticas diferentes.

La *IM* calcula la diferencia entre la entropía (medida de la incertidumbre) de la variable dependiente y la entropía de la misma variable cuando se conoce el valor de la variable independiente. En esta métrica que tiene un rango de cero a uno, valores cercanos a cero indican independencia entre las variables, mientras que valores cercanos a uno marcan una alta correlación entre ellas. En la Ec. (2), se ilustra el cálculo como la diferencia de la entropía *H* de la variable *X* y la entropía de esa variable cuando se conoce el valor de una variable *Y*:

$$I(X:Y) = H(X) - H\left(\frac{X}{Y}\right) \quad (2)$$

La selección de las longitudes de onda a ser incluidas en cada uno de los modelos de regresión se realizó de tres maneras diferentes. La primera fue incluyendo las primeras 20 longitudes de onda con mayor *FM* y las primeras 20 con mayor *IM* en un solo set de calibración; la segunda fue identificando regiones de interés de forma visual al graficar el 10% de las variables con mayor *IM* y el 10% con mayor *FM* de forma independiente; y la tercera consistió en utilizar un algoritmo que seleccionara el set de longitudes de onda que maximizara la *IM*, de tal forma que ésta no aumentaría agregando ninguna de las longitudes de onda restantes.

2.d. Regresión y validación de los modelos

Después de aplicar los tres tipos de selección de variables mencionados anteriormente a cada uno de los sets de datos que se tenían para cada tipo de matriz (2 de agua en MIR, 2 de plasma en MIR, 2 de sangre en MIR, 1 de agua en NIR, 2 de plasma en NIR, y 2 de sangre en NIR), se llevaron a cabo tres tipos de regresión: Regresión lineal simple (RLS); Regresión de mínimos cuadrados (PLSR) y regresión de componentes principales (PCR). El esquema de la selección de variables y los modelos de regresión aplicado a cada set de datos, se ilustra en la Fig. 2.

Se obtuvo entonces un total de 324 modelos de regresión para la región MIR, y 270 modelos para la región NIR, a los cuales se calculó el error

Fig. 2: Esquema de identificación de longitudes de onda y aplicación de modelos de regresión.

cuadrático medio de validación cruzada (RMSECV por sus siglas en inglés: *Root Mean Square Error of Cross Validation*), como indicador de la capacidad de predicción del modelo, debido a que el procedimiento “*leave one out*” de la validación cruzada permite evaluar todo el rango de concentraciones y trabajar con modelos contruidos a partir de una cantidad adecuada de muestras en el set de calibración.

3. Resultados

Para cada uno de los set de datos se calcularon los índices *IM* y *FM*, y se obtuvieron gráficos como los que se muestran en las Figs. 3 y 4, respectivamente.

Se seleccionaron los 15 modelos con menor RMSECV de cada tipo de muestra, y se obtuvieron las longitudes de onda que hacían parte de cada uno de ellos. A partir de esta clasificación inicial, se identificaron las técnicas de procesamiento con mejores resultados, se reportaron los errores RMSECV correspondientes, y se elaboró un resumen que se muestra en las Tablas I y II con las combinaciones de las técnicas de procesamiento, en el cual los espacios vacíos indican que para la muestra de esa columna, el procesamiento de esa fila no fue seleccionado dentro del grupo de modelos con mejor comportamiento de predicción.

Estas tablas se realizaron para cada región, y se llevó a cabo la selección de los cinco modelos más significativos por región de acuerdo a su aparición en la mayoría de sets de muestras

(sombreados), los cinco con menor promedio total (subrayados) y se identificó de forma adicional los modelos con menor RMSECV para las muestras DGA, PSM y SNG de forma individual (valores en negrilla). Se identificó una marcada repetición de longitudes de onda en los modelos seleccionados, entre 648 cm^{-1} y 1250 cm^{-1} y entre 3000 cm^{-1} y 3870 cm^{-1} , común tanto para las muestras de agua como para las de plasma y sangre.

Con poca aparición en las muestras de agua, pero con presencia significativa en muestras de plasma y sangre, se identificaron longitudes de onda entre 1850 cm^{-1} y 2210 cm^{-1} , y entre 2350 cm^{-1} y 2565 cm^{-1} . De igual forma se identifican, aunque con menor repetición, algunas longitudes entre 1380 cm^{-1} y 1726 cm^{-1} para dichas muestras.

Por otro lado, en la región NIR se identifican longitudes de onda entre 11100 cm^{-1} y 11950 cm^{-1} , y entre 5190 cm^{-1} y 5700 cm^{-1} en la mayoría de los modelos seleccionados, tanto para las muestras de agua, como para las de plasma y sangre. Uno de los modelos que utiliza la selección del set principal seleccionado a partir de la Información Mutua, presenta de forma aislada longitudes de 5958 cm^{-1} y 5890 cm^{-1} , cercanas a la región anterior.

Se observa la presencia de longitudes entre 8790 cm^{-1} y 9265 cm^{-1} en una mayor proporción para muestras de plasma y sangre que para muestras de agua, mientras que longitudes entre 4300 cm^{-1} y 4900 cm^{-1} se presentan en varios de los modelos de agua, y en menor proporción en las muestras de plasma.

Longitudes entre 6000 cm^{-1} y 6760 cm^{-1} se tuvieron presencia exclusivamente en las muestras de plasma de varios de los modelos, y se evidencia una menor cantidad de longitudes entre 10400 cm^{-1} y 10900 cm^{-1} tanto en muestras de plasma como de sangre.

Con el fin de identificar de forma más clara regiones potencialmente útiles en la predicción de glucosa en las matrices estudiadas, se evaluó la frecuencia de aparición de las longitudes seleccionadas en las regiones anteriores tanto en el espectro NIR como en el MIR.

Fig. 3: Factor de mérito de muestras de plasma en la región del cercano infrarrojo (NIR): Señales filtradas sin procesamiento adicional correspondientes a muestras del sujeto A.

Fig. 4: Información mutua de en la región del cercano infrarrojo (NIR): Señales filtradas sin procesamiento adicional correspondientes a muestras del sujeto A.

Como resultado se obtuvo una mayor importancia para longitudes entre 648 cm^{-1} y 1250 cm^{-1} y entre 2350 cm^{-1} y 2565 cm^{-1} para el espectro MIR, mientras que en el NIR sobresalieron longitudes entre 11100 cm^{-1} y 11950 cm^{-1} , y entre 5190 cm^{-1} y 5700 cm^{-1} . Estas regiones se muestran de forma gráfica en las Figs. 5 y 6.

4. Discusión

De los 18 modelos seleccionados que incluyen las regiones MIR y NIR, 12 correspondían al

procesamiento derivativo, mientras que otros 6 correspondían a procesamiento de corrección polinomial de línea base, cálculo de variable normal estándar y señales originales. Lo anterior indicando los cálculos de primera y segunda derivada como procesamientos óptimos en la identificación de las longitudes de onda que presentan relación con la concentración de glucosa en las muestras.

Se identificó también que 13 de los modelos seleccionados utilizaron las principales longitudes de onda teniendo en cuenta los

TABLA I
Selección de modelos en la región del medio infrarrojo (MIR).

Técnicas Utilizadas	RMSECV REGIÓN MIR									
	DGA1	DGA2	PSMA	PSMB	SNGA	SNGB	PROM DGA	PROM PSM	PROM SNG	PROM TOTAL
DI, PRIN, PLS	27.0	39.6	31.8	26.9	71.4	62.2	33.3	29.4	66.8	43.2
DII, PRIN, PCA	34.6	34.1	54.7	25.0	95.9	60.7	34.4	39.9	78.3	50.8
DI, PRIN, PCA	20.2	28.1	33.9	25.1	92.0		24.2	29.5	92.0	39.9
DII, PRIN, PLS	32.6	28.3	57.6	35.9		70.1	30.5	46.8	70.1	44.9
DI, IM, PLS		48.0	35.8	26.9	49.6	74.1	48.0	31.4	61.9	46.9
ORIGINAL, IM, PLS		26.2	55.4				26.2	55.4		40.8
ORIGINAL, PRIN, PLS			54.8	28.6				41.7		41.7
ORIGINAL, PRIN, PCA			58.1	28.6				43.4		43.4
POLINOMIAL, PRIN, PCA				25.6		77.7		25.6	77.7	51.7

TABLA II
Selección de modelos en la región del cercano infrarrojo (NIR).

	RMSECV REGIÓN NIR								
	DGA1	PSMA	PSMB	SNGA	SNGB	PROM DGA	PROM PSM	PROM SNG	PROM TOTAL
DI, IM, PLS	56.9	29.4	68.3	56.2	104.6	56.9	48.9	80.4	63.1
DII, PRN, PLS	65.8	27.2	73.4	77.6	106.8	65.8	50.3	92.2	70.2
DII, IM, PLS	45.3	20.4		61.4	56.8	45.3	20.4	59.1	46.0
DI, PRN, PLS	57.8	48.0	43.4		47.9	57.8	45.7	47.9	49.3
DII, PRN, PCA	73.5	31.1	75.6	72.9		73.5	53.4	72.9	63.3
DI, PRN, PCA	48.8		36.4		49.8	48.8	36.4	49.8	45.0
SNV, PRN, PLS			45.0	45.6			45.0	45.6	45.3
DI, REG, PLS	24.6	69.6				24.6	69.6		47.1

criterios de información mutua y factor de mérito en conjunto, 4 utilizaron la selección exclusiva del set de datos que maximiza la información mutua, y tan solo 1 utilizó una región como fuente de datos. De esta forma se hace evidente la utilidad de combinar la ponderación de ambas métricas en la selección de las longitudes de onda óptimas para la predicción.

En las Figs. 3 y 4 se evidencia la congruencia entre la ponderación que realiza la *IM* y el *FM*. Sin embargo, la utilización de ambas métricas en simultáneo permitió una mejor identificación que se ve evidenciada en que 13 de los 18 modelos con menor RMSECV utilizaron la combinación de estas como criterio de selección de variables a incluir en los modelos de regresión.

Los hallazgos presentados coinciden parcialmente con diferentes autores que han trabajado de forma individual las regiones NIR y MIR. En las Tablas III y IV se identifican las regiones y longitudes de onda coincidentes con

estudios anteriores, y se identifican regiones no especificadas anteriormente. De igual forma se subrayan las regiones identificadas como de mayor importancia dentro de los resultados del estudio, y se realiza un énfasis con un cuadro alrededor de las regiones que se identificaron en mayor proporción en las muestras de plasma y sangre.

Estas últimas deberían ser analizadas de forma más detallada con el fin de identificar si corresponden a compuestos exclusivos del plasma y la sangre que no se relacionan con los niveles de glucosa en dichas matrices.

Lo anterior permite ver que para la región MIR únicamente uno de los intervalos de longitud de onda identificados como de mayor importancia (subrayados) se encuentra presente en una alta proporción en los modelos de plasma y sangre, mientras que para la región NIR, ninguno de los mismos está dentro de esta selección, y para ninguna de las dos regiones existe una coincidencia con los intervalos reportados en la literatura.

Fig. 5: Identificación de regiones más relevantes en el espectro Infrarrojo Medio (MIR): Señales filtradas sin procesamiento adicional correspondientes a muestras del sujeto A.

Fig. 6: Identificación de regiones más relevantes en el espectro Infrarrojo Cercano (NIR): Señales filtradas sin procesamiento adicional correspondientes a muestras del sujeto A.

Se evidencia que las técnicas de selección de variables utilizadas no habían sido aplicadas por los autores que se relacionan en las Tablas III y IV, lo cual permite corroborar la validez de las mismas en la identificación de variables de predicción para la glucosa al coincidir con algunos de los hallazgos anteriores.

Los errores RMSECV promedio obtenidos para la región MIR en las diferentes matrices oscila entre 24 mg/dl y 62 mg/dl. Se identifica el aumento drástico del error al trabajar con una muestra de mayor complejidad como lo es la sangre, en la cual se obtuvo un valor de 61,9

mg/dl como menor error RMSECV, mientras que para las muestras de agua y plasma se obtuvieron errores hasta de 24,2 mg/dl y 25,6 mg/dl respectivamente.

Para la región NIR, a pesar de haber obtenido errores de 45,3 mg/dl y 20,4 mg/dl para las muestras de agua y plasma respectivamente, el menor error reportado para las muestras de sangre fue de 47,9 mg/dl, lo cual evidencia un mejor desempeño de esta región en comparación con la MIR para muestras de alta complejidad.

TABLA III
Regiones coincidentes con estudios anteriores en la región cercana del infrarrojo (NIR).

Región hallada en Cercano Infrarrojo (NIR)	Región Coincidente	Autor
<u>5958 cm⁻¹ y 5890 cm⁻¹</u>	5924 cm ⁻¹	Dr. Pulenta: Armónico de la vibración C-H (2005) [9]
<u>5190 cm⁻¹ a 5700 cm⁻¹</u> <u>11100 cm⁻¹ a 11950 cm⁻¹</u>	5296 cm ⁻¹ y 7194 cm ⁻¹	Y-J- Kim. G. Yoon (2006) [20]
	N/A	N/A
8790 cm⁻¹ a 9265 cm⁻¹	N/A	N/A
4300 cm ⁻¹ a 4900 cm ⁻¹	4180 cm ⁻¹ a 4892 cm ⁻¹	Y-J- Kim. G. Yoon (2006) [20]
6000 cm ⁻¹ a 6760 cm ⁻¹ ; 10 400 cm ⁻¹ a 10900 cm ⁻¹	N/A	N/A

TABLA IV
Regiones coincidentes con estudios anteriores en la región media del infrarrojo (MIR).

Región hallada en Medio Infrarrojo (MIR)	Región Coincidente	Autor
<u>648 cm⁻¹ y 1250 cm⁻¹</u>	1000 cm ⁻¹ a 1230 cm ⁻¹ ; pico en 1033 cm ⁻¹	Dr. Petibois: Absorción enlace C-O (1999) [21]
	1022 cm ⁻¹ a 1119 cm ⁻¹	Y-J- Kim, S. Hahn, y G. Yoon. (2003) [22]
	1032 cm ⁻¹ , 1042 cm ⁻¹ , y 993 cm ⁻¹	Vargas y colaboradores (2009 y 2011) [7, 9]
	800 cm ⁻¹ a 1275 cm ⁻¹	Absorción teórica enlace C-O-C
<u>1380 cm⁻¹ y 1726 cm⁻¹</u> <u>1850 cm⁻¹ a 2210 cm⁻¹ ; 2350 cm⁻¹ y 2565 cm⁻¹</u>	N/A	N/A
3000 cm ⁻¹ a 3870 cm ⁻¹	3120 cm ⁻¹ a 3570 cm ⁻¹	Absorción teórica enlace O-H
	3020 cm ⁻¹ a 3085 cm ⁻¹	Absorción teórica enlace C-H

A pesar de haber obtenido errores relativamente bajos para ambas regiones, algunos estudios que aplican la Información Mutua como técnica de selección de variables [18] reportan un mejor desempeño al emplear modelos de regresión no lineales posterior a la selección por IM, por lo cual una posible razón para no haber obtenido un rango menor del error RMSECV pudo obedecer a la utilización de modelos de regresión lineales como lo son LR, PLSR, y PCR.

5. Conclusión

Los resultados de la presente investigación se prestan para realizar diferentes análisis y contrastes estadísticos debido a las combinaciones de técnicas aplicadas. Estas combinaciones conjuntas de métodos de procesamiento no se realizaron debido a que la aplicación de más de un método de pretratamiento espectral puede llevar a la eliminación de información útil [16]; sin embargo, los resultados del estudio permitirían

iniciar pruebas aplicando las técnicas identificadas como más acertadas de forma individual.

La selección de los mejores modelos de regresión de acuerdo al criterio del RMSECV, permitió concluir que tanto la primera como la segunda derivada realizan una mejor identificación de las longitudes de onda que se relacionan con la concentración de glucosa y brindan una alta información acerca de la misma.

En relación a las técnicas de selección utilizadas, se concluyó que la mezcla entre las variables más relevantes de acuerdo a la *IM* y el *FM* presentaron una mayor correlación con la concentración de glucosa que otro tipo de selección como la que comprende del set que maximiza la *IM*, o las regiones identificadas de forma visual. Al obtener este resultado con la técnica mencionada, se hace evidente que la implementación de un algoritmo similar al utilizado para el caso de *IM* sería de gran utilidad en estudios futuros.

El algoritmo de selección de variables por el método que maximiza la *IM* se hizo presente en 4 de los modelos seleccionados, por lo que también debería ser incluido en estudios futuros. Sin embargo, el método de identificación visual de regiones de interés no es recomendable como método de selección de variables.

Esto puede conllevar a estudios enfocados en alguna de las regiones utilizadas para la construcción de los modelos y una validación más exhaustiva de los mismos, al igual que la utilización de modelos no lineales de predicción.

De acuerdo a las Tablas III y IV, de las longitudes de onda subrayadas como variables más importantes, se observa una coincidencia con varios autores para la región MIR comprendida entre 648 cm^{-1} y 1250 cm^{-1} , y para la región NIR entre 5190 cm^{-1} y 5700 cm^{-1} .

Sin embargo, también se identificó que algunas de estas regiones no se encuentra reportadas en la literatura consultada, como lo son las longitudes de onda entre 2350 cm^{-1} y 2565 cm^{-1} para el espectro MIR, y entre 11100 cm^{-1} y 11950 cm^{-1} para el espectro NIR.

El hecho de que las regiones de mayor importancia seleccionadas en todo el estudio tengan poca coincidencia con las más frecuentes en los modelos de plasma y sangre evidencia la necesidad de más estudios enfocados en muestras de alta complejidad, las cuales pueden incluir también simuladores de tejido con el fin de tener un mayor acercamiento a la medición no invasiva de la molécula.

En relación a los valores RMSECV obtenidos, se debe resaltar que valores desde $20,4\text{ mg/dl}$ en plasma, hasta $61,9\text{ mg/dl}$ en sangre, son errores que no serían aceptados en el ámbito clínico, por lo cual el método tal y como se describe debe ser mejorado con técnicas de procesamiento de señales para ser lo suficientemente fiable y permitir una medición que pueda estar cerca de implementarse en el ámbito médico.

Agradecimientos

Los autores agradecen a la Dra. Liliana Tirado y la laboratorista Juliana Jaramillo por su colaboración en la utilización del equipo de medición espectroscópica en el Instituto Interdisciplinario de las ciencias de la Universidad del Quindío. Este trabajo es patrocinado por la Universidad Autónoma de Occidente (UAO) bajo el proyecto No. 10INTER-132 y por el Departamento Administrativo de Ciencia, Tecnología e Innovación Colciencias por el soporte al Joven Investigador John Edward Vargas.