



INAOE

**DETECCIÓN DE BANDAS DE PROTEÍNAS DE
INTERÉS BIOLÓGICO EN GELES DE
POLIACRILAMIDA**

Presentada por

Jorge Jaime Juárez Lucero

Como requisito parcial para alcanzar el grado de

**Doctorado en Ciencias y Tecnologías
Biomédicas**

En el

**INSTITUTO NACIONAL DE ASTROFÍSICA, ÓPTICA Y
ELECTRÓNICA.**

Enero, 2023,

Tonantzintla, Puebla

Directores de tesis:

Dr. Leopoldo Altamirano Robles

Dra. Anabel Socorro Sánchez Sánchez

©INAOE 2023

Derechos Reservados

El autor otorga al INAOE el permiso de reproducir y distribuir copias de esta tesis en su totalidad o en partes mencionando la fuente.



Resumen

La sobre expresión de la proteína GPN está relacionada con el incremento del cáncer de mama ductal y lobular invasivo (CDI y CLI) tipo HER2+, se ha propuesto su uso como biomarcador para emplearla como método de diagnóstico de la enfermedad. Sin embargo, los análisis solamente pueden realizarse por técnicas moleculares. El método más simple para detectar la presencia de proteínas es buscarlas en un gel de poliacrilamida (SDS-PAGE), pero las técnicas actuales aplicadas a las imágenes de los geles no permiten detectar las muestras ni las proteínas presentes sin ayuda del experimentador. Tampoco existen bases de datos de imágenes de geles de GPN que permitan identificar sus diferentes niveles de sobre expresión, por lo que se desarrolló una metodología que utilizó las diferentes fuerzas químicas que existen en las proteínas y mantuvo la GPN soluble y con una elevada pureza que permitió emular dichas muestras para conseguir las imágenes de geles SDS-PAGE.

Se propuso una nueva metodología denominada Perfil de Imagen Basado en Segmentación de Imágenes Binarias (PIBSIB) que, empleando una máscara binaria, de tamaño de 1x400 pixeles para estudiar muestras o de tamaño de 1x50 pixeles si se buscan proteínas, binarizó y obtuvo el valor de intensidad del pixel de la posición 255, después de recorrer toda la imagen generó un arreglo que al ser graficado produjo un nuevo perfil de imagen. Los mínimos de este nuevo perfil se relacionaron con el número de muestras presentes y los máximos detectaron la cantidad de proteínas presentes, con ello, de forma automática se encontró el peso molecular de las proteínas junto con la de menor y mayor sobre expresión en las diferentes muestras.

Abstract

The GPN protein over expression is related with increase of breast cancer invasive ductal and invasive lobular (IDC and ILC) HER2+ type, Its use as a biomarker has been proposed as a method of diagnosis of the cancer disease. However, analysis can only be performed by molecular techniques. The simplest method to detect the presence of proteins is to look for them using a polyacrylamide gel (SDS-PAGE), but current techniques applied to gel images do not allow detection of the samples or the proteins present without the help of the experimenter. There are also no databases of GPN gel images that make it possible to identify their different levels of overexpression, so a methodology was developed that used different chemical forces that exist in proteins interactions and kept the GPN soluble with a high purity that allowed to emulate samples of people with cancer disease to obtain the images of SDS-PAGE gels.

A new methodology called Image Profile Based on Segmentation of Binarized Images (IPBSBI) was proposed, this methodology to use a binary mask, with a size of 1x400 pixels to study samples or a size of 1x50 pixels if proteins are sought, binarized and obtained the value of intensity of the pixel at position 255, after going through the entire image, generated an arrangement that plotted, produced a new image profile. The minimums of this new profile were related to the number of samples present into the gel and the maximums detected the number of proteins present in each sample, with this, the molecular weight of the proteins was automatically found together with the lowest and highest overexpression in the different samples.

AGRADECIMIENTOS

Agradezco al CONAHCYT y al Instituto Nacional de Astrofísica, Óptica y Electrónica por el apoyo recibido al proporcionarme beca, software e instalaciones que me ayudaron a la realización de esta investigación doctoral. Así como al Instituto de Ciencias de la BUAP por la realización de los experimentos para la obtención de los geles de poliacrilamida.

DEDICATORIAS

Este trabajo doctoral va dedicado con todo mi cariño para mi familia:

Mi esposa María del Rayo por su confianza, amor y tiempo que me permitió realizar los experimentos tanto de laboratorio como todo el desarrollo del trabajo.

A mi hija Aileen, por su amor, apoyo y por nuestros desvelos juntos en la realización de mi trabajo.

A mi hijita Georgy, por su amor y su apoyo al ser mi consejera espiritual.

Contenido

Resumen	2
Abstract.....	3
Capítulo 1. Introducción	9
1.2. Problema para resolver	16
1.3. Pregunta de investigación.....	17
1.4. Solución propuesta.....	17
1.5. Hipótesis	18
1.6. Objetivo general.....	18
1.7. Objetivos específicos	18
1.8. Aportaciones	19
1.9. Publicaciones resultantes de esta investigación	20
1.10. Delimitaciones.....	22
1.11. Estructura de la tesis	23
Capítulo 2. Estado del arte	24
2.1. Electroforesis de geles de ADN y proteínas en una y dos dimensiones	24
2.2. Técnica IMAC de purificación de proteínas recombinantes	26
2.3. Técnicas de imágenes para los análisis de geles de poliacrilamida y DNA en 1 y 2 dimensiones	29
2.4. Segmentación semántica para la detección de objetos.....	37
Capítulo 3. Solución Propuesta	42
3.1. Emulación de geles de pacientes con diferente sobre expresión de proteína.....	42
3.1.1. Construcción del vector de expresión para la GPN	44
3.1.2. Preparación de la columna His-bind.....	45
3.1.3. Expresión de la proteína GPN y crecimiento a diferentes temperaturas ..	45
3.1.4. Preparación a diferentes concentraciones de buffer con NaCl	48
3.1.5. Preparación a diferentes concentraciones de buffer con TRIS	48
3.1.6. Preparación a diferentes concentraciones de buffer con EDTA	48
3.1.7. Preparación a diferentes concentraciones de buffer con DTT	49
3.1.8. Purificación con aminoácidos a diferentes concentraciones	49
3.1.9. Purificación con Buffer Final	49

3.1.10. Cuantificación de la proteína GPN.....	50
3.1.11. Análisis de dispersión de luz dinámica.....	50
3.1.12 Emulación de muestras con diferentes niveles de sobre expresión de la proteína GPN para representar distintos estadios de cáncer de mama relacionados con la sobre expresión de la proteína.....	51
3.2. Desarrollo de algoritmos de análisis de imágenes y segmentación semántica.....	52
3.2.1. Adquisición de la imagen.....	54
3.2.2. Preprocesamiento y extracción de características empleando una nueva metodología denominada perfil de imagen basada en segmentación de imágenes binarias (PIBSIB).....	54
3.2.3. Extracción de características empleando segmentación semántica.....	58
Capítulo 4. Resultados obtenidos.....	60
4.1. Construcción del vector de expresión para la GPN y prueba de expresión....	60
4.2. Preparación de la columna His-bind.....	61
4.3. Expresión de la proteína GPN y crecimiento a diferentes temperaturas.....	61
4.4. Análisis con diferentes concentraciones de NaCl.....	63
4.5. Análisis con diferentes concentraciones de TRIS.....	64
4.6. Análisis con diferentes concentraciones de EDTA.....	66
4.7. Análisis con diferentes concentraciones de DTT.....	67
4.8. Purificación con aminoácidos.....	67
4.9. Purificación con Buffer Final.....	70
4.10. Cuantificación de la proteína GPN.....	71
4.11 Análisis de dispersión de luz dinámica.....	72
4.12. Análisis preliminar de los geles de proteína GPN a diferentes concentraciones.....	73
4.13 Preprocesamiento y Extracción de características empleando metodología desarrollada (perfil de imagen basado en segmentación de imágenes binarias empleando una máscara binaria).....	75
4.14. Detección de proteína sobre expresada en geles SDS-PAGE empleando Perfil de Imagen basado en segmentación de imágenes binarias.....	78
Capítulo 5. Discusión, conclusiones y trabajo a futuro.....	106
5.1. Discusión.....	106
5.2. Conclusiones.....	115

5.3. Trabajo futuro.....	119
Apéndices.....	120
A. Lista de Figuras.....	120
B. Lista de tablas.....	123
Referencias	124
Resumen en inglés	135

Capítulo 1. Introducción

Los geles de poliacrilamida o de sacarosa se han empleado dentro de la biomedicina y biología, para buscar la ausencia o expresión de segmentos de genes de ADN o de proteínas de interés biológico, por lo general relacionados con algún padecimiento o enfermedad. El estudio de estos geles de proteínas dentro del análisis de imágenes es un área de oportunidad de aplicación. Recientemente técnicas de procesamiento de imágenes han sido empleadas para mejorar la calidad de los geles, tanto de ácidos nucleicos como de proteínas en una y dos dimensiones. La mayoría de las técnicas se han basado en eliminar el ruido de fondo que permita realizar una detección semi-automática de alguna zona o región de interés, empleando el perfil de la imagen para mejorar la expresión de las manchas de proteínas.

La búsqueda de la presencia, ausencia o sobre expresión de proteínas de interés biológico relacionadas con diferentes padecimientos o enfermedades, se realiza por medio de técnicas de laboratorio que pueden ser muy caras como el empleo de HPLC (*High-Performance Liquid Chromatography* por sus siglas en inglés) y otras no tan costosas como el SDS-PAGE (*sodium dodecyl sulfate polyacrylamide gel electrophoresis* por sus siglas en inglés), *Immunoblotting*, *ELISA*, *CD Spectroscopy*, *transmission electron microscopy* y *LC/MS* entre otras [1]. En el área biomédica la alteración de la información genética o la expresión incorrecta de proteínas puede llevar a un desequilibrio en la salud, conduciendo a diferentes tipos de enfermedades congénitas, leves, graves o agudas [2], que en el peor de los casos puede provocar una muerte prematura. En estos casos un pronto diagnóstico ayuda a prevenir el desarrollo de una enfermedad grave e incluso la muerte.

Los geles de poliacrilamida se han empleado como un método de detección para encontrar fragmentos de genes [3, 4] o proteínas específicas para diagnosticar enfermedades [5, 6, 7]. Algunos de los genes buscados en los

geles de electroforesis detectan la presencia de VIH, hepatitis C, influenza H1N1, COVID-19 y permiten distinguir entre diferentes tipos de dengue [8, 9, 10, 11, 12]. En los geles de proteínas se ha realizado la detección de antígeno prostático específico, elevadas concentraciones proteicas relacionadas con cáncer de mama, cáncer de colon y Alzheimer como lo muestra la figura 1 [13, 14, 15, 16], esta característica los hace ideales como un método de diagnóstico de enfermedades [5, 6, 7].

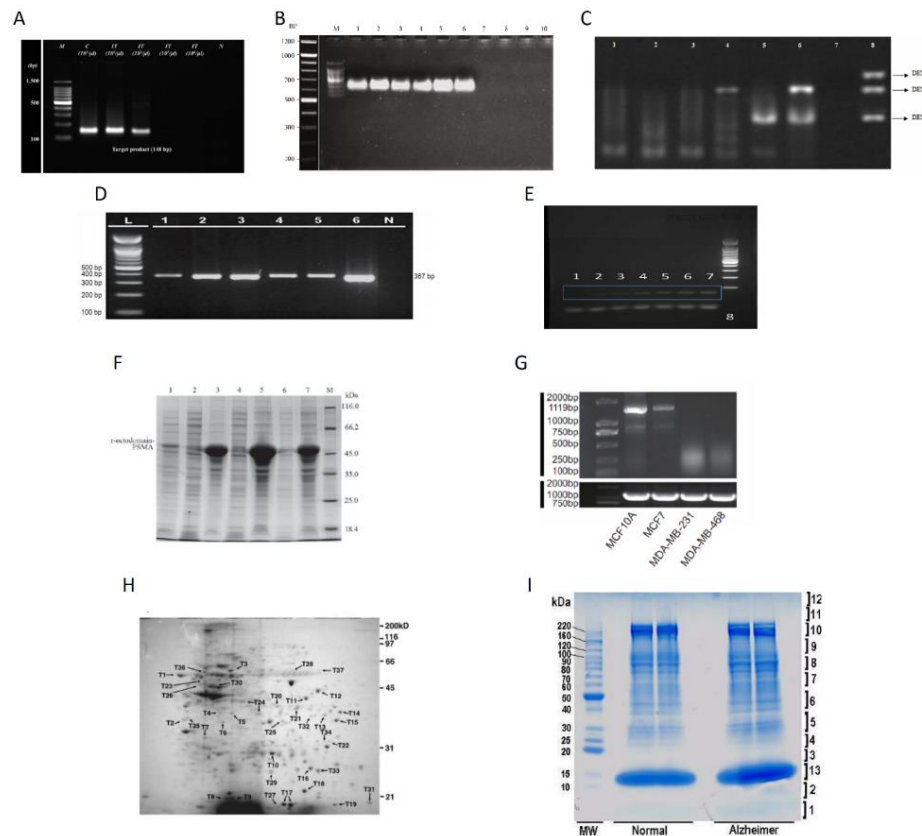


FIGURA 1. GELES DE ELECTROFORESIS DE GENES.

A) Los tres primeros pacientes presentan infección de VIH y los dos últimos se encuentran sanos (Fuente: [8]). B) Seis pacientes con hepatitis C (Fuente: [9]). C) Paciente 8 con tres cepas de virus de dengue (Fuente: [10]). D) Diferente nivel de expresión del virus H1N1 en 6 pacientes (Fuente: [11]). E) Siete pacientes con diferente concentración de virus SARS-CoV-2. Geles de electroforesis de proteínas (Fuente: [12]). F) Personas con distintas concentraciones de PSMA provocando diferente grado de cáncer de próstata (Fuente: [13]). G) Detección de diferentes proteínas relacionadas con el cáncer de mama. (Fuente: [14]) H) Proteínas presentes en el cáncer de colon (Fuente: [15]). I) Expresión de altas concentraciones de proteínas relacionadas con la enfermedad de Alzheimer (Fuente: [16]).

De todas las afecciones, el cáncer es considerado de mayor importancia, es la causa principal de muerte en el mundo. En 2018 se reportaron más de 18 millones de casos nuevos junto con 9 millones de muertes. Se espera que para el 2040 se incremente hasta 29 millones de enfermos con más de 16 millones de muertes [17]. Por lo anterior, los médicos buscan realizar un diagnóstico que permita tomar mejores decisiones a favor de los pacientes.

En particular, el cáncer de mama ocupa el primer lugar a nivel internacional en incidencias y el quinto en defunciones (ver figura 2), es la primera causa de muerte de mujeres y se presenta como una afección heterogénea con múltiples variaciones morfológicas y moleculares [18, 19].

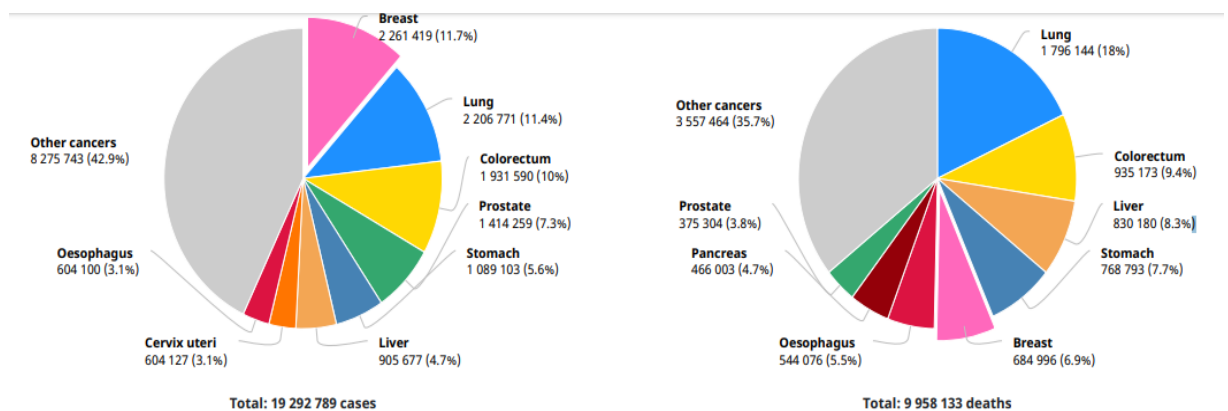


FIGURA 2. INCIDENCIAS Y DEFUNCIONES DE CÁNCER DE MAMA A NIVEL INTERNACIONAL.
Fuente: [19].

En México, el cáncer de mama se ha incrementado de manera alarmante afectando a mujeres de 20 años cubriendo un rango de 15 defunciones por cada 100 mil mujeres [18]. América del Norte contiene el 89.4% de afecciones por cada 100,000 personas de las cuales el 12.5% resulta ser de tipo mortal, ver figura 3 [19].

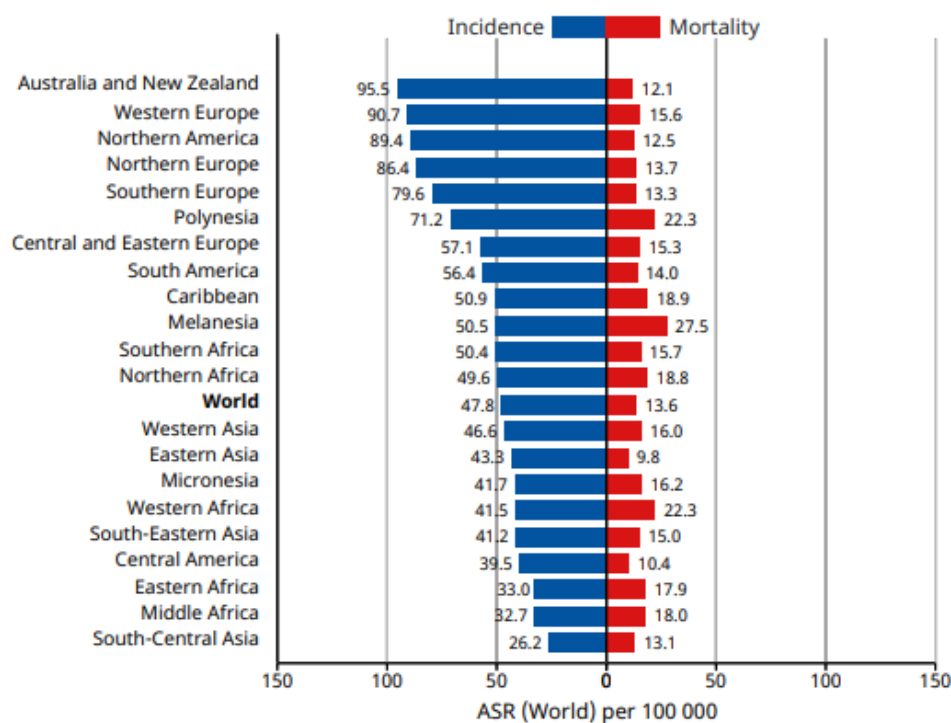


FIGURA 3. FLUJO DE INCIDENCIA Y MORTALIDAD DEL CÁNCER DE MAMA A NIVEL INTERNACIONAL. Fuente: [19].

Se han detectado 6 diferentes tipos de cáncer de mama: Bajo en claudina, tipo basal, Her2+, similar a la mama normal, luminal A y luminal B, de los cuales el 15% de las afecciones pertenecen al Her2+, 40% al luminal A y 20% al luminal B donde estos tres últimos cubren el 75% de los casos de cáncer mamario y son altamente invasivos catalogados como el CDI (Carcinoma Ductal Invasivo) y el CLI (Carcinoma Lobular Invasivo). El CDI es el tipo más común ya que 8 de 10 enfermos lo presentan y afecta los ductos de leche materna (Figura 4) y se vuelve invasivo al resto de los tejidos empleando el sistema linfático y el torrente sanguíneo para la metástasis. Sin embargo, el CLI (Figura 4) es menos común ya que lo presentan 1 de 10 enfermos, inicia en los lóbulos mamaros antes de entrar en metástasis y es muy difícil de detectar ya sea con análisis físicos, mamografía o estudiando las imágenes de los tejidos, siendo

su invasión de tal grado que afecta ambos senos y solamente es detectable cuando el cáncer se encuentra en etapa avanzada [20, 21, 22, 23, 7].

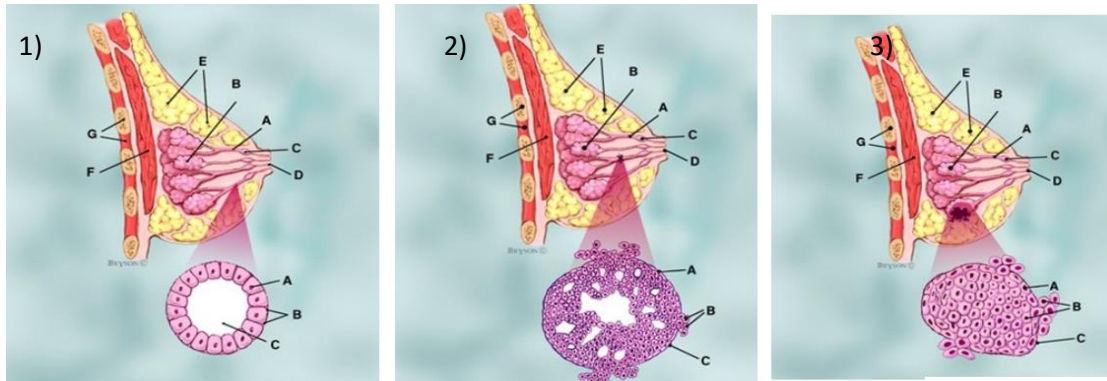


FIGURA 4. COMPARACIÓN ENTRE CÉLULAS NORMALES Y AFECTADAS.

1) Representación de las células mamarias normales, 2) células mamarias con CDI, 3) células mamarias con CLI. Imagen superior: A= Ductos, B= Lóbulos, C= Ductos de leche, D= pezón, E= grasa, F= músculos del pecho, G= Costillas. Imagen inferior: A= Células que recubren el lóbulo, B= Células que rompen la membrana basal. C= Membrana basal. Fuente: www.breastcancer.org.

Los métodos de diagnóstico de cáncer de mama son: exploración de las mamas (detectable cuando existen tumores), mamografía (presenta células cancerígenas dentro de los conductos mamarios como manchas blancas), biopsias (cirugía menor para recuperar tejido para realizar un estudio clínico). Estas técnicas resultan caras, requieren tiempo de análisis que puede llevar en el mejor de los casos de tres días hasta una semana para conocer los resultados, y en ciertos tipos de cáncer solamente son detectados cuando ya se tiene la metástasis en estado avanzado [20].

Por otro lado, a nivel molecular se han encontrado algunas proteínas GTPasas involucradas en el desarrollo de cáncer, algunas como Ras han presentado altos niveles de concentración en diferentes tipos de cáncer favoreciendo la producción de oncogenes [24], Rho está estrechamente vinculada con la iniciación de la metástasis [25], Rab27 promueve la progresión de melanomas [26]. Sin embargo, de todas las GTPasas conocidas la sobreexpresión de la proteína GPN ha estado vinculada de forma significativa con los tipos de

carcinomas ductal invasivos ER+, HER2+, y el luminal tipo A y B. En el caso de los CDI ER+ la sobre expresión de la proteína GPN reduce el porcentaje de supervivencia hasta un 40 % lo que se traduce en una esperanza de vida de 5000 días para los enfermos; por su parte en el caso del CDI HER2+ la sobre expresión de la proteína reduce la supervivencia del paciente hasta un 0% permitiendo un máximo de 1,500 días de vida o aproximadamente 4 años, ver figura 5 [27].

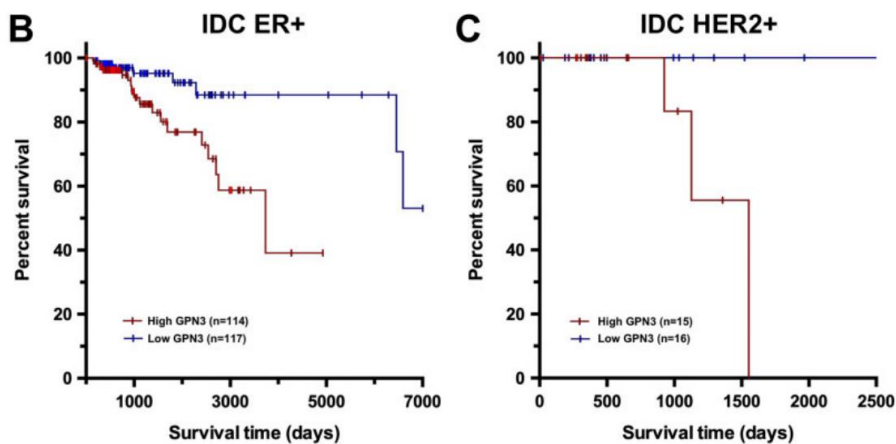


FIGURA 5. SUPERVIVENCIA RELACIONADA CON SOBRE EXPRESIÓN DE PROTEÍNA GPN.

Efecto en la supervivencia de pacientes con cáncer tipo CDI ER+ y HER2+ al tener sobre expresión de la proteína GPN. Fuente: [27].

Esta relación existente entre la presencia de cáncer CDI ER+ y HER2+ con la sobre expresión de la proteína GPN, ha sugerido la búsqueda de la sobre expresión de todas las GTPasas relacionadas con los diferentes tipos de cáncer mamario y emplearlas como biomarcadores de la presencia de diferentes tipos de cánceres [25, 24, 27, 21]. Las técnicas propuestas para detectar la sobreexpresión de las GTPasas, incluyendo la GPN; son por medio de técnicas moleculares como PCR, secuenciación de genes o purificación de proteínas.

La presencia de proteínas en un extracto celular puede detectarse a menor costo y sin el empleo de técnicas moleculares elaboradas, ya que solamente se requiere de lisar las células de forma mecánica con morteros por sonicación o empleando detergentes; homogenizarla y centrifugarla. Al sobrenadante se le realiza la electroforesis SDS-PAGE y se revela el gel empleando azul de coomasie [28]. Sin embargo, existen factores extrínsecos que distorsionan las muestras o las proteínas y que afectan la interpretación que se le dé al gel, como pueden ser: 1) Destrucción de la muestra por causa del tipo de agarosa que se utilice (figura 6A), 2) Exceso de proteína cargado por muestra (líneas 3, 4 y 5 de figura 6B), 3) Flujo de migración lento que no permite la separación de proteínas (líneas 4 y 7 de figura 6C), 4) Rayas generadas por proteína precipitada (línea 4 de figura 6D), 5) Rayas provocadas por cantidades desproporcionadas de proteínas (líneas 4 y 10 de figura 6E), 6) Gel curvado por cambios en el voltaje empleado como lo muestra la figura 6F, [29, 30, 7].

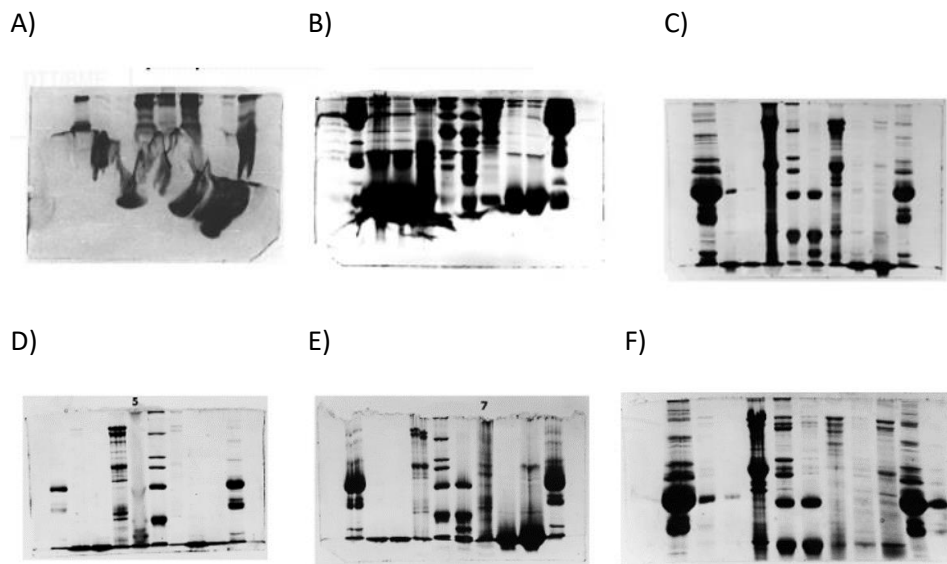


FIGURA 6. GELES SDS-PAGE DAÑADOS.

Geles de poliacrilamida con diferentes afectaciones que deforman las proteínas y/o las muestras presentes en el gel dificultando su interpretación. Fuente: <https://www.ruf.rice.edu/~bioslabs/studies/sds-page/sdsgoofs.html>.

A pesar de que un gran número de enfermedades pueden ser diagnosticadas con el empleo de geles, y que el cáncer de mama invasivo se encuentra estrechamente ligado con la sobre expresión de algunas proteínas, incluyendo la GPN; existe aún la necesidad de desarrollar tecnología que permita la identificación de estos componentes biológicos de forma semi-automática o automática corrigiendo los factores intrínsecos que pueden distorsionar una muestra. Esta acción puede ser realizada mediante el empleo de computadoras y desarrollo de software en la práctica clínica, para ayudar a diagnosticar enfermedades y para facilitar un tratamiento adecuado a los pacientes, recientemente una gran cantidad de técnicas utilizan herramientas de *machine learning*, para desarrollar clasificadores que permitan analizar los datos biomédicos; facilitando la identificación y el diagnóstico de enfermedades para mejorar la calidad de vida de quienes las padecen [31, 32]. Lo anterior, con el fin de disminuir o eliminar incluso el error humano, permitiendo detectar la presencia y sobre expresión de proteínas relacionadas particularmente con cáncer de mama dentro de la imagen del gel de poliacrilamida, de tal modo que sea posible buscar la sobre expresión proteica sin depender solamente de técnicas moleculares.

1.2. Problema para resolver

Detectar de forma semiautomática en imágenes de geles de poliacrilamida la ausencia, presencia o sobreexpresión de una proteína específica, relacionada con el diagnóstico de una enfermedad causada por su sobreexpresión. Como caso de estudio se analizará la sobreexpresión de la proteína GPN humana purificada, que se encuentra relacionada con el cáncer de mama invasivo ductal y lobular tipo Luminal Her2+.

1.3. Pregunta de investigación

¿En qué medida los algoritmos de análisis de imágenes y de segmentación semántica tienen la capacidad de detectar la presencia y sobreexpresión de la proteína GPN relacionada con cáncer invasivo de mama ductal y lobular tipo luminal Her2+ que además permitan desarrollar un sistema confiable que pueda servir como una posible herramienta de diagnóstico y sea aplicado a las imágenes de geles de poliacrilamida de las diferentes proteínas de interés médico o biológico?

1.4. Solución propuesta

La solución propuesta para resolver el problema consiste en los siguientes pasos:

- 1) Emplear segmentación semántica para eliminar el ruido de fondo.
- 2) Aplicar una máscara binaria que genera un arreglo que contiene solamente los valores del píxel blanco, para generar una gráfica que representa un nuevo perfil de la imagen.
- 3) Emplear los mínimos del nuevo perfil para identificar el número de muestras presentes en el gel.
- 4) Emplear los máximos del nuevo perfil aplicado a la muestra para identificar el número de bandas de proteínas presentes.
- 5) Identificar el valor de los máximos para relacionarlos con la concentración o cantidad de proteína GPN expresada o sobre expresada por muestra.

1.5. Hipótesis

La aplicación de algoritmos de segmentación semántica en imágenes de geles de proteínas permitirá identificar y cuantificar la presencia y sobreexpresión de la proteína GPN relacionada con cáncer de mama, presente en las muestras analizadas.

1.6. Objetivo general

Desarrollar una técnica empleando algoritmos de análisis de imágenes y segmentación semántica, para detectar la sobre expresión de una proteína relacionada con cáncer invasivo ductal y lobular tipo luminal Her2+, utilizando imágenes de geles de poliacrilamida SDS-PAGE, y verificar su uso para detectar sobre expresiones de diferentes tipos de células cancerígenas.

1.7. Objetivos específicos

1. Producción de geles SDS-PAGE empleando extractos celulares de proteínas para la emulación de:
 - a) Un mínimo de 200 muestras de pacientes con presencia de proteínas GPN en cantidades normales.
 - b) Un mínimo de 200 muestras de pacientes con la presencia de proteínas GPN sobre expresadas.
2. Generar una base de imágenes que contenga geles SDS-PAGE de la proteína GPN en forma no expresada y expresada.

3. Desarrollar un algoritmo para eliminación del ruido de fondo que facilite la segmentación y permita separar las columnas y bandas de un gel de poliacrilamida de proteínas.
4. Desarrollar una red neuronal que realice segmentación semántica que posibilite la detección de la sobre expresión de las proteínas GPN en geles de poliacrilamida de proteínas.
5. Empleo y cuantificación de la metodología desarrollada para la detección de la sobre expresión de diferentes tipos de células relacionadas con distintos padecimientos de cáncer.

1.8. Aportaciones

Como resultado de la investigación doctoral se desarrolló una metodología denominada: Perfil de Imagen Basado en Segmentación de Imágenes Binarias. Esta metodología está compuesta de los siguientes puntos:

- 1) Un algoritmo que permite la eliminación de ruido y mejora la presentación de los geles de poliacrilamida para poder detectar la presencia de proteínas con base a su peso molecular.
- 2) Un algoritmo que permite detectar la cantidad de líneas que corresponden a diferentes muestras y detecta las bandas de proteínas que corresponden a cada experimento.
- 3) Desarrollo de una CNN que realiza segmentación semántica para clasificar el gel SDS-PAGE y eliminar el ruido de fondo.

Cada uno de los puntos anteriores produjo una contribución importante de la investigación doctoral.

1.9. Publicaciones resultantes de esta investigación

Artículos en Journal (JCR):

Development of a Methodology to Adapt an Equilibrium Buffer/Wash Applied to the Purification of hGPN2 Protein Expressed in Escherichia coli Using an IMAC Immobilized Metal Affinity Chromatography System. Jorge Juárez-Lucero, María del Rayo Guevara-Villa, Anabel Sánchez-Sánchez, Raquel Díaz-Hernández, and Leopoldo Altamirano-Robles. *Separations*, 2022, 9, 164, pp. 1-16.

Image Profile Based in Binarized Image Segmentation to detect over expression of GPN protein related to breast cancer IDC and ILC Her2+ type. Artículo en preparación.

Capítulos de libro:

Collaborative Learning to Teach Set Theory to Engineer Students. Jorge Juárez, María del Rayo Guevara-Villa, Anabel Sánchez, Raquel Díaz, Leopoldo Altamirano. *Transactions on Computational Science and Computational Intelligence.* Springer. Electronic ISSN 2569-7080

Image Segmentation Applied to Line Separation and Determination of GPN2 Protein Overexpression for Its Detection in Polyacrylamide Gels. Jorge Juárez, María del Rayo Guevara-Villa, Anabel Sánchez, Raquel Díaz, Leopoldo Altamirano. *Lecture Notes in Computer Science. Progress in Artificial Intelligence and Pattern Recognition.* Springer. ISSN 0302-9743

Tridimensional Structure Prediction Purification of Human Protein GPN2 to High Concentrations by Nickel Affinity Chromatography in Presence of Amino Acids for Improving Impurities Elimination. Jorge Juárez, María del Rayo Guevara-Villa, Anabel Sánchez, Raquel Díaz, Leopoldo Altamirano.

Transactions on Computational Science and Computational Intelligence.
Springer. Electronic ISSN 2569-7080

Artículos en conferencias:

Methodology to Predict Unknown Protein Structure. GPN Protein Family a Case of Study. J. J. Juárez, A. S. Sánchez, R. Díaz, J. Newton, M. R. G. Guevara-Villa and L. Altamirano, 2021 *Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE)*, 2021, pp. 1-5, doi: 10.1109/GMEPE/PAHCE50215.2021.9434836.

Application of an Artificial Neural Network to Classify countries with COVID-19 adjusting to SIR model. Juárez, J., Sánchez, A., Díaz, R., Altamirano, L. *Memorias del congreso Nacional de Ingeniería Biomédica.* (2020). Pp. 454-461. Recuperado de <http://memorias.somib.org.mx/index.php/memorias/article/view/797>

Identificación de Bandas de Proteína Recombinante Expresada en *Escherichia coli* Usando una máscara binaria. Jorge Juárez, María del Rayo Guevara Villa, Anabel Sánchez, Raquel Díaz, Leopoldo Altamirano. *Enfrentando retos emergentes de Ciencia y Tecnología.* (2020). ISBN 978-958-771-986-4. Pp. 113-117.

KNIME para la comparación de diferentes clasificadores del modelo SIR de países infectados por COVID-19. Aileen Juárez, María del Rayo Guevara, Anabel Sánchez, Nicolas Galeno Razo, Jorge Juárez. *Congreso Internacional México – Colombia CICOM 2020.* (2020). ISSN 2462-9588. Pp. 206-211.

Artículos en revistas nacionales:

ANOVA to compare three methods to track COVID-19 in nine countries. J. J. Juárez-Lucero, A. S. Sánchez-Sánchez, R. Díaz-Hernández, M. R. Guevara-Villa, L. Altamirano-Robles. *Revista Mexicana de Ingeniería Biomédica.* (2021). ISSN 2395-9126. Pp. 36-46.

Presentación de póster:

Desarrollo de geles de Poliacrilamida para simular la sobre expresión de proteína relacionada con cáncer de mama. V Seminario Internacional en Ciencias y Tecnologías Biomédicas – PRIS 2022.

Los genes KIR, cáncer de mama e infecciones virales. IV Seminario Internacional de Ciencias y Tecnologías Biomédicas, PRIS-BIOMÉDICAS 2021.

Patentes:

PURIFICACIÓN DE PROTEÍNAS RECOMBINANTES MEDIANTE CINÉTICAS DE AGREGADOS A UN BUFFER MAESTRO. INVENTORES: Jorge Jaime Juárez Lucero, Anabel Socorro Sánchez Sánchez, Raquel Díaz Hernández, María del Rayo Graciela Guevara Villa, Leopoldo Altamirano Robles. Publicada en la Gaceta de la Propiedad Industrial. SOLICITUDES DE PATENTE, DE REGISTROS DE MODELO DE UTILIDAD Y DE DISEÑOS INDUSTRIALES. Enero. 2023. MX/a/2022/006275. Aprobado el examen de forma.

1.10. Delimitaciones

El presente trabajo cuenta con las siguientes limitaciones:

1. Sólo se cuentan con 382 geles con 10 muestras cada uno.
2. Algunos procesos de software requieren tener la licencia de MATLAB

1.11. Estructura de la tesis

El primer capítulo explica la motivación de la investigación doctoral incluyendo la introducción al tema, el planteamiento del problema a resolver, la pregunta de investigación, la solución propuesta, la hipótesis de investigación, los objetivos, aportaciones, las publicaciones resultantes y las delimitaciones. El capítulo 2 incluye el estado del arte que permite comprender la investigación doctoral. El capítulo 3 presenta la solución propuesta para resolver el problema planteado. El capítulo 4 muestra todos los resultados obtenidos de la investigación doctoral. El capítulo 5 analiza las discusiones de la tesis, se detallan las conclusiones y el trabajo a futuro.

Capítulo 2. Estado del arte

El siguiente capítulo está enfocado a comprender las bases teóricas de esta investigación doctoral. En la primera parte, se explican los tipos de geles de poliacrilamida de uno y dos dimensiones y la técnica empleada para conseguir las proteínas puras. En la segunda sección, se describen las técnicas de análisis de imágenes empleadas para extraer información de los geles de diferentes dimensiones y finalmente, se analiza el empleo de la segmentación semántica para lograr la clasificación de objetos.

2.1. Electroforesis de geles de ADN y proteínas en una y dos dimensiones

La técnica de electroforesis es usada para lograr la separación de proteínas y secuencias de genes de ADN, aplicando una carga eléctrica sobre las muestras para que puedan desplazarse por el gel de agarosa o poliacrilamida, de tal modo que puedan ser retenidas en alguna posición dentro del gel, a causa de su tamaño lo que impide su completa migración y pérdida; debido a las cargas eléctricas que presentan las moléculas.

La distribución de las proteínas o ADN dentro de un gel en una dimensión consiste en un arreglo de columnas y filas. Las columnas representan diferentes condiciones del experimento o diferente número de pacientes que estén en análisis. Las filas dentro de cada columna contienen un número finito de bandas donde cada una de ellas representa una proteína específica, la cual está contenida en cada extracto celular de cada una de las muestras. Las bandas más gruesas indican una concentración mayor de la proteína dentro de la muestra [33, 34]. A diferencia de estos, los geles en dos dimensiones se llevan a cabo realizando dos separaciones, la primera con base a su punto isoeléctrico y la segunda tomando en cuenta su peso molecular (ver figura 7).

Los estudios de biomedicina molecular en el área de proteómica dependen de la interpretación que se dé a los geles de poliacrilamida. Interpretación ligada a la calidad que tenga el gel, lo cual se muestra en la nitidez de la imagen obtenida. Una imagen confusa del gel de la expresión de proteínas puede generar conclusiones o diagnósticos incorrectos. Por ejemplo: al analizarlo varios expertos, pueden etiquetar distintas bandas como si fuera la misma. También puede ocurrir el efecto contrario, que distintos expertos etiqueten la misma banda como si se tratara de diferentes proteínas. Estos eventos, pueden generar falsos positivos o falsos negativos, pudiendo diagnosticar en el mejor de los casos a una persona sana como enferma y en el peor de los casos a un enfermo como sano. En estos casos, es necesario complementar el diagnóstico con estudios clínicos o moleculares en el paciente. Además, esto provoca que el analista tenga que repetir experimentos para poder verificar si una proteína se ha expresado o no. Esto conlleva en el peor de los casos a tener que solicitar nuevamente al paciente la toma de muestras para poder buscar alguna proteína de interés biomédico, incrementando los costos y tiempo para el diagnóstico tanto para el paciente como para el hospital; mientras la enfermedad puede ir avanzando a etapas peligrosas que ponga en riesgo la vida del paciente.

Por lo regular los geles de ADN representan una dificultad menor para su interpretación porque cada línea presenta una sola banda del gen expresado (figura 1, A-E). En cambio, los geles de proteínas que pueden incluir un número elevado de las mismas poseen un mayor grado de dificultad para su interpretación. Esto considerando que el ojo humano disminuye su capacidad para detectar diferencias entre dos intensidades similares o iguales, por lo que es probable que se generen errores de interpretación al analizar un gel de proteínas, ya sea por ilusiones ópticas, visión sensible o fatiga [7, 35, 36, 37, 38].

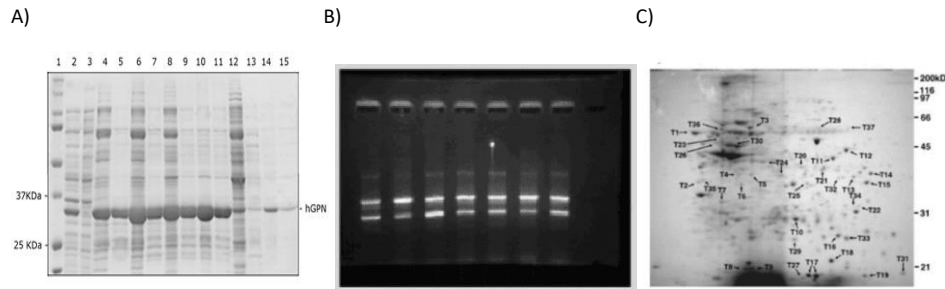


FIGURA 7. DIFERENTES TIPOS DE GELES.

(A) Expresión y purificación de la proteína GPN en un gel SDS-PAGE, la columna 1 representa al marcador de peso molecular y las columnas 2 a la 15 son diferentes experimentos realizados. Las filas representan las proteínas expresadas en cada uno de los experimentos. (B) Gel de agarosa de ADN, primera y última columna representa el marcador de peso molecular, columnas 2 a la 6 son experimentos diferentes, las filas indican el gen expresado. (C). Gel en dos dimensiones de una sola muestra en una sola columna para facilitar la identificación de proteínas. Fuente: A) Elaboración propia, B) [33], C) [15].

Las muestras que incluyen las proteínas analizadas en los geles de poliacrilamida pueden ser obtenidas directamente de extractos celulares, lo que hace esta técnica una metodología eficiente para identificar alguna proteína de interés biomédico. Además, sin descartar que puede ser utilizada para verificar el grado de pureza de alguna proteína específica, ya sea al emplear distintos tipos de buffers o usando diferentes técnicas de purificación. La técnica más económica para purificar proteínas es la cromatografía de afinidad de metales inmovilizados (IMAC por sus siglas en inglés, *Immobilized Metal Affinity Chromatography*) que se analizará a continuación.

2.2. Técnica IMAC de purificación de proteínas recombinantes

La purificación de proteínas por medio de un sistema IMAC requiere que se inserte el gen de la proteína específica que se quiere expresar dentro de un vector o plásmido de tipo pET28, junto con una secuencia de seis aminoácidos histidinas (hexahistidinas), esta proteína de interés se le conocerá como proteína recombinante.

El plásmido es insertado dentro de la bacteria *Escherichia coli* (*E. coli*) empleando choque térmico, es decir, se colocan los vectores de expresión (plásmidos) en la solución que contiene las bacterias en un tubo Eppendorf, se introduce por tres minutos a 37 °C en un baño termostático y se enfría durante 2 minutos, este proceso finalmente permite que el gen que incluye la proteína recombinante pueda introducirse dentro de la bacteria (ver figura 8).

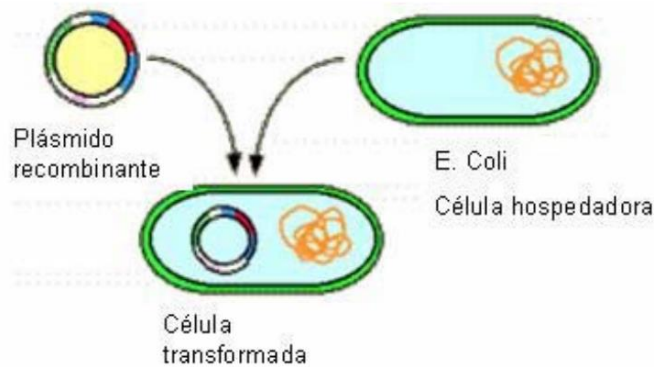


FIGURA 8. BACTERIA MODIFICADA.

Inserción de vector de expresión dentro de la bacteria *E. coli* por medio de choque térmico. Fuente: [39].

Se hace crecer la bacteria en un medio de cultivo y se utiliza en el proceso de purificación. Se lisa la bacteria para recuperar el contenido (las proteínas nativas junto con la proteína recombinante) y se prepara la columna siguiendo las indicaciones del fabricante.

La columna contiene dentro de la resina un metal (níquel) que será afín a la hexahistina que se incluye en la proteína recombinante, se le introduce la muestra y se deja rotando para permitir que la proteína de interés pueda adherirse a la resina, como lo muestra la figura 9. Posteriormente, la columna es lavada con diferentes tipos de soluciones amortiguadoras para ir eliminando los contaminantes (las proteínas que no pudieron adherirse en la columna). Finalmente, se eluye con un buffer a base de imidazole que compite por los

enlaces presentes entre la hexahistidina y el níquel, para liberar la proteína de interés que se espera se encuentre completamente en forma pura con el mínimo número de contaminantes (ver figura 10).

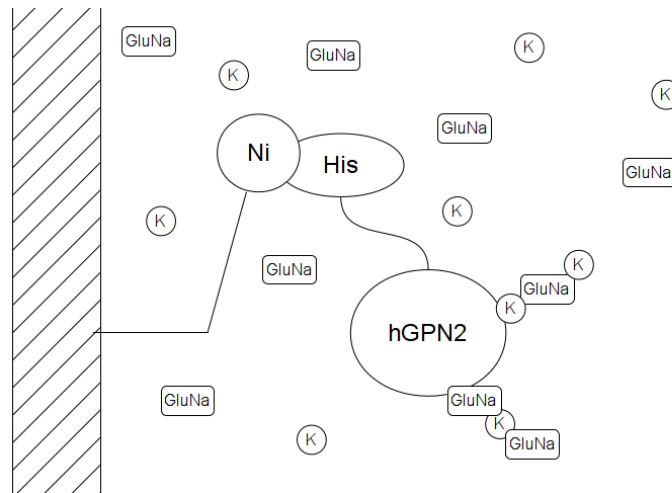


FIGURA 9. ENLACE PROTEÍNA-COLUMNA.

Proteína recombinante con hexahistidinas adherida a la columna por medio de los enlaces presentes entre el níquel de la resina y el aminoácido de la proteína. Fuente: Elaboración propia.

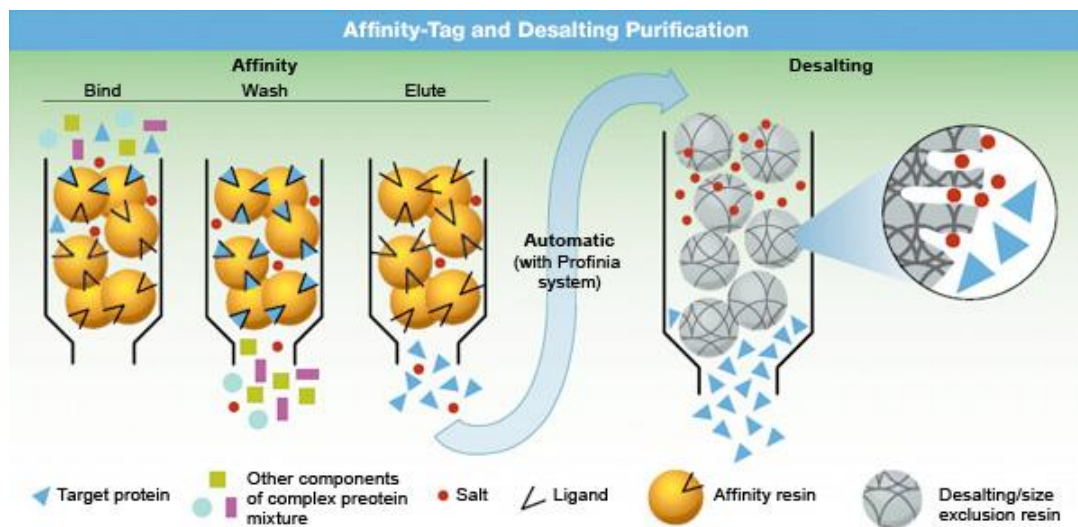


FIGURA 10. COLUMNA IMAC.

Columna de purificación de proteínas empleando cromatografía de afinidad por metales inmovilizados junto con una columna de desalado proporcionado por el kit de Bio-Rad. Fuente: <https://www.bio-rad.com/>.

Finalmente, para cambiar el buffer con contenido elevado de imidazole y dejar la proteína recombinante dentro de una solución amortiguadora que permita realizar sus funciones metabólicas, se emplea una columna de desalado y la proteína pura ya puede ser utilizada en los estudios pertinentes o almacenada a -80°C como lo muestra la figura 10 [40, 41].

La pureza de la proteína estudiada se comprueba empleando geles de poliacrilamida o SDS-PAGE (*sodium dodecyl sulfate–polyacrylamide gel electrophoresis*). Debido a la importancia de los geles para la detección de proteínas y tomando en cuenta el creciente error humano al ser analizados (fatiga del investigador, errores de interpretación, mal criterio de búsqueda, desarrollo de geles de mala calidad, bandas mal interpretadas, falta de experiencia del analista, efectos ópticos) se han realizado diversas técnicas para que, empleando análisis de imágenes, se puedan procesar los geles de poliacrilamida como se explica en el siguiente capítulo.

2.3. Técnicas de imágenes para los análisis de geles de poliacrilamida y DNA en 1 y 2 dimensiones

En 1999 se publicó por parte del equipo de Xiangyun Ye el primer acercamiento sobre como procesar un gel de poliacrilamida empleando transformadas “*Top Hat*”, para eliminar el fondo de la imagen seguido de un análisis de picos del histograma de la imagen para poder detectar la existencia o ausencia de bandas en una región específica del gel [35]. Para el 2000, Alon Efrat plantea el uso de algoritmos geométricos para fusionar bandas que se encuentren cerca, para puedan ser identificadas como una sola de tal modo que puedan ser comparadas con algún gel de referencia. Reconoce que el principal problema en el análisis de imágenes de geles de proteínas es identificar las bandas, para hacer una correlación de la misma banda en dos geles diferentes de 2D donde las bandas corren a lo largo y ancho [42]. Ivan

Bajla en el 2005 presenta un análisis de geles de ADN eliminando ruidos de fondo empleando transformadas de Fourier a diferentes regiones del gel, que deben seleccionarse por parte del usuario para ajustar las bandas a una rejilla de acuerdo con su peso molecular [43]. Kaabouch en 2007 emplea umbralización automática para igualar los contornos a grises de la imagen de fondo y aplica filtros para mejorar la calidad de la imagen de los geles, lo cual ayuda al analista a decidir cuales resultados experimentales son confiables, ya que durante el desarrollo de la electroforesis se pueden producir geles con la calidad suficiente para minimizar los errores de interpretación [33]. Labyed identifica que los pasos principales para analizar un gel de electroforesis requieren de la detección de la banda, correlacionarla a su peso molecular, cuantificarla y compararla. Labyed propone el cálculo de la desviación estándar de las intensidades de los pixeles de las columnas junto con sus derivadas, para identificar los máximos donde se encuentran las bandas de las proteínas con el requisito de tener las columnas bien identificadas [44]. Ramaswamy en 2010 busca mejorar la calidad de la imagen de los geles utilizando un valor de umbral para normalizar y eliminar el ruido. Las proteínas empleadas en el gel son almacenadas en una base de datos, siendo el usuario quien elige la región de la imagen que se quiere analizar [45].

Para 2011 Tsakanikas emplea transformadas contourlet para fusionar bandas en geles 2D e identificar una misma proteína, cuando en el gel ocurren corrimientos de la muestra eliminando los pixeles cercanos y partiendo la banda que corresponde a la misma proteína. También divide la imagen en regiones para aplicar transformadas gaussianas para encontrar los puntos máximos y eliminar el resto considerándolo como ruido de fondo [46].

Por su parte, Jiann-Der Lee reconoce que el análisis de los geles por parte del especialista puede llevar a muchos errores donde se ignore información importante por lo que diseña un procedimiento automático para geles de ADN usando procesamiento digital de imágenes para eliminar el fondo y algoritmos

fuzzy c-means para segmentar la imagen reagrupándola con el uso de funciones gaussianas que permitan estimar las bandas del gel [36]. Y Soto plantea el empleo de la función *downhill simplex* junto con algoritmos genéticos, para encontrar mínimos locales en los puntos donde se encuentran las bandas y para optimizarlos se busca su centro utilizando funciones gaussianas para cada mínimo local [47].

Taher en el 2013 realiza mejoras a los geles para evitar la repetición de experimentos empleando dos filtros a la imagen, uno espacial y otro de frecuencia (filtro gaussiano pasa bajas) sobre las intensidades de la imagen [37] mientras que Koprowski analiza los diferentes cambios que sufren de brillo en diferentes regiones del gel para utilizar cada cambio como separador entre bandas [38].

En 2014 Magdeldin realiza una comparación de los programas comerciales realizados por diferentes empresas de biomedicina y biotecnología demostrando que la mayoría solamente realiza técnicas simples de preprocesamiento como el manejo de brillo o contraste, y en algunos casos son capaces de dividir las regiones de interés [48]. Brauner propone un nuevo algoritmo basado en calcular el área que ocupan los píxeles relacionados con las bandas, para cuantificar la cantidad de proteína en el gel habiendo previamente elegido las bandas por el usuario haciendo ajustes con suavizados gaussianos para remover el ruido [49]. En ese mismo año Taher publica el cálculo del umbral óptimo para poder eliminar posibles bandas falsas basándose en el perfil de intensidad de la imagen para poder detectar las bandas verdaderas [37].

Abadi realiza correcciones a los geles que presentan el efecto de carita feliz, con ayuda del usuario puede detectar líneas que contienen los espacios que separan las columnas, emplea filtros para eliminar el ruido de fondo hasta que casi desaparezca la imagen, dejando solamente rastros de las bandas. A esta nueva imagen se resta la original para detectar el ruido, resta nuevamente el

ruido de la imagen original hasta visualizar solamente las proteínas y detectarlas, midiendo el perfil de intensidad con el defecto de perder bandas importantes dentro del gel [30]. Por su parte, Abeykoon empleando geles de ADN los cambia a grises redimensionándolos aplicando filtros de mediana para minimizar el ruido, crea histogramas evaluando negros y blancos sacando promedios eligiendo el valor 100 como umbral para separar bandas del fondo; como los máximos no pueden ser detectados por presentar múltiples picos en cada columna encuentra su posición midiendo la variación de grises y luego utiliza el mismo procedimiento para detectar las bandas calculando su peso molecular utilizando el buffer de carga junto con una distribución exponencial [50]. Intarapanich genera un programa denominado GELect donde consigue separar líneas distorsionadas con efecto de carita sonriente, las regiones donde se encuentran las bandas de interés son recortadas por el usuario para que el software pueda medir el perfil de intensidad y encontrar los máximos evaluando derivadas de primer orden para localizar las bandas de proteínas específicas [4]. Sim utiliza GelApp que emplea filtros Gabor para segmentar los geles usando solamente aquellos que presenten muy pocas proteínas, que no se encuentren traslapadas y sean fácilmente detectables [51].

Para el 2016 Rezaei divide el gel en pequeñas regiones y emplea la densidad espectral mediante un análisis de Fourier para calcular automáticamente el ancho de banda promedio de la línea, elimina el ruido con transformadas *wavelet* y filtros de mediana; su método detecta máximos para las líneas y manualmente se elige el umbral específico que permita identificar todas las columnas presentes [52].

Por su parte, Fernández-Lozano identifica texturas empleando transformadas *wavelets* y de Fourier; manualmente se eligen las proteínas y las bandas falsas para aplicar diferentes métodos de clasificación [53]. Turan busca contornos empleando el algoritmo de *Canny* y redes neuronales artificiales para definir las clases de bajo y alto umbral (proporcionado por el usuario) y compara sus

resultados con los detectores de bordes *Sobel*, *Prewitt* y *Robert* [54]. El grupo de Verbeek emplea técnicas de procesamiento digital de imágenes para encontrar un valor de umbral que es restado de la imagen original, proporcionando un histograma mejor distribuido utilizando una máscara para hallar el valor máximo que permita restablecer los valles vecinos de cada pico. Para detectar las bandas buscan el área y perímetro de cada una [34].

En 2018, Goez hace una revisión de las técnicas empleadas como herramientas de preprocesamiento para eliminar ruido de fondo en geles de 2D entre las que se destacan cambios de intensidad, umbralización, filtros lineales, suavizado, filtros gaussianos, transformadas *contourlet* o *wavelet*. Para buscar las bandas o manchas de proteínas reporta que se han realizado ajustes polinomiales junto con la búsqueda de cambios en los mínimos locales y modificaciones en el histograma [7].

Alnamoly desarrolla el programa *EGBioImage* que requiere de correcciones manuales de las bandas cuando se encuentran muy cercanas, emplea regresión polinomial para hallar el peso molecular de proteínas y en caso de no detectarlo es necesario que sea proporcionado por el usuario. Se tiene que elegir la región de interés donde se encuentra la banda deseada para utilizar *K-means* como clasificador de las bandas que pertenezcan al mismo peso para verificar que se encuentran en la misma posición. Para la detección de las bandas o proteínas emplea el teorema de Green [29].

Ou propone el uso del algoritmo de *watersheds* en geles 2D para poder identificar si la mancha corresponde a una sola proteína o dos [55]. Finalmente, esta tesis propone el empleo de máscaras binarias para elegir regiones pequeñas de los geles y aislar una muestra con el objetivo buscar de forma automática un umbral, que permita ir segmentando las proteínas presentes de tal forma que mediante un contador de bandas, se pueda ir definiendo el umbral que permite detectar solamente proteínas sobre expresadas [56].

Los programas actuales desarrollados para filtrar las bandas y eliminar ruido de fondo (Scanalytics, GelcomparII, Gel-Pro Analyzer, TotalLab, PDQuest, Proteomweaver, Dcyder 2D, imageMAster, Melanie, BioNumerics, Redfin, Gel IQ, Z3, Delta2D Flicker, EGBioImage, GelCluster, GelAnalyzer, PyElph, GelQuant y QuantiScan) presentan las desventajas de ofrecer un procesamiento parcial de imágenes con características muy limitadas y algunos de ellos (Gel Plugin, ImajeJ, GelAnalyzer, GelClust, GelQuant.NET, Image, Laneruler y PyElph) son catalogados como software libre con funciones limitadas. La mayoría no puede detectar las columnas, ni mucho menos las bandas, ni proporcionar el peso molecular de alguna proteína específica. Para el caso de los softwares que sí pueden detectar columnas y bandas requieren de ayuda del usuario y la gran mayoría está realizado para analizar geles de ADN o 2D. En el mejor de los casos solamente tienen una precisión del 54% en la detección de bandas [4] con un error de hasta el 17% reportado por Abeykon [50]. Otros programas resultan demasiado complicados para que el analista pueda utilizarlos y no todos permiten la agrupación de elementos comunes [33, 29].

En la tabla 1, se presentan las características principales de algunos de los programas desarrollados por los autores que han realizado investigación, para detección de bandas en geles de ADN y de proteínas en una o dos dimensiones. En general, todas las investigaciones realizadas en el análisis de imágenes de geles de ADN y proteínas buscan conseguir tres procesos: Eliminar ruido de fondo (*background*), detectar líneas y segmentar las bandas.

Para eliminar el ruido de fondo se han empleado diferentes técnicas entre las que se incluyen transformadas *contourlet* y *wavelet*, transformadas *Top-Hat*, filtros pasa baja Gaussianos, normalización, cambios en la intensidad, filtros de mediana, umbral adaptativo, gaussianos no lineales, análisis de Fourier, *fuzzy-c-means* junto con alguna matriz de convolución, cortes en la imagen buscando regiones de interés. Estos métodos se han empleado por separado

o combinados entre sí con la finalidad de conseguir perfiles de la imagen mejorados, libres de ruido para poder extraer las características del gel [33, 34, 4, 7, 35, 36, 38, 57, 30, 50, 43, 49, 42, 53, 58, 44, 48, 45].

Para detectar las líneas y/o las bandas del gel se han aplicado herramientas como; elección de píxeles clave, para relacionar el cambio de contraste con el fondo buscando una mejor segmentación; detectar bordes con aproximaciones bayesianas junto con métodos de umbralización o segmentación de Otsu; elección de región de interés (ROI) por parte del usuario, para reducir tamaño y minimizar ruido para calcular la desviación estándar, eligiendo un umbral de forma manual y generar el perfil del gel. Otras herramientas emplean funciones gaussianas o uso de plantillas, que permitan dividir el gel en zonas ROI que contengan las líneas especificadas previamente por el usuario; aplicación de filtros Sobel, que detectan los máximos y mínimos en los perfiles de intensidad para identificar los espacios entre las líneas o muestras; medir cambios en el brillo, contar el número de píxeles que pertenecen a cada una de las muestras para medir sus distancias empleando variaciones en los niveles de grises; calcular la densidad espectral, para promediar el ancho de las líneas lo que permite elegir regiones específicas dentro de la imagen. Otros analizan el perfil de los picos por medio de sus áreas, para agruparlos empleando técnicas como *K-means*. Una opción más delimita las bandas con elipses buscando que no ocurra intersección entre ellas para encontrar la separación de las líneas. De todos los métodos, los que mejores resultados han descritos incluyen la elección de zonas ROI por parte del usuario, para delimitar ruido y generar mejores perfiles que permitan encontrar los mínimos que se relacionan con la separación de las líneas que contiene el gel [33, 34, 4, 7, 35, 36, 38, 29, 56, 30, 43, 49, 42, 53, 57, 44, 45, 52].

TABLA 1. TABLA COMPARATIVA.

Programas y avances de los trabajos desarrollados en el área de análisis de imágenes de geles de ADN y proteínas. Fuente: Elaboración propia.

Autor/SW	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
PyElph, 2012	N	N	N	N	N	N	N	N	N	-	-	-	-
GelCluster, 2013	S	S	N	N	N	N	N	N	N	-	-	-	-
Sim, 2015	N	S	N	N	N	N	N	N	N	N	S	S	S
Abadi, 2015	S	S	-	S	N	N	N	N	N	N	S	S	S
Intarapanich, 2015	S	S	S	N	N	N	N	N	S	N	S	N	S
Turan, 2016	N	S	N	N	N	N	N	N	N	N	S	N	S
Fernández-Lozano, 2016	N	N	N	N	N	N	N	N	S	N	N	S	N
Rezaei, 2016	S	N	N	N	N	N	N	N	S	N	S	N	S
Verbeek, 2017	N	S	-	N	S	N	N	N	N	N	S	S	N
Goez, 2018	N	N	S	N	N	N	N	N	N	N	N	S	N
Ou, 2020	N	S	S	N	N	N	N	N	N	N	N	S	N
Alnamoly (EGBioImage), 2020	S	S	S	S	S	N	N	N	N	S	S	N	S
GelAnalyzer, 2022	S	N	N	S	N	N	N	N	N	-	-	-	-

C1: Detección de columnas
 C2: Detección de bandas
 C3: Requiere el usuario colocar el nombre de las muestras
 C4: Proporciona el peso molecular de forma automática
 C5: Realiza la agrupación de las bandas empleando *K-means*
 C6: Usa el buffer de carga para calcular el peso molecular
 C7: Usa la imagen del gel para encontrar las proteínas de baja expresión
 C8: Usa la imagen del gel para encontrar proteínas sobre expresadas
 C9: Separa regiones concentradas de proteínas
 C10: Emplea métodos de interpolación para encontrar el peso molecular
 C11: Es utilizado para analizar geles de una dimensión
 C12: Es utilizado para analizar geles de proteínas
 C13: Es utilizado para analizar geles de ADN
 S: Sí
 N: No

En consideración de que la mayoría de los métodos desarrollados para realizar los análisis de imágenes de geles, se basan en mejorar los perfiles de la imagen para detectar tanto las muestras presentes como las bandas de proteínas de cada línea. En este trabajo doctoral se propone el empleo de otras técnicas, que permitan la identificación tanto de las líneas que corresponden a las muestras como de cada una de las proteínas presentes en cada una de ellas; como puede ser la segmentación semántica que se menciona en el apartado siguiente.

2.4. Segmentación semántica para la detección de objetos

La segmentación semántica permite agrupar píxeles de una imagen y acomodarlos en clases que pueden relacionarse con diferentes objetos, así se puede crear una semántica para cada uno de los diferentes objetos que contiene la imagen como puede ser personas, calles, semáforos, edificios, árboles, etc.

La ventaja de esta técnica es que permite encontrar los contornos de cada figura que se localizan dentro de la imagen una vez que se ha definido la clase, entonces si ya se ha etiquetado la clase “EDIFICIO”, la segmentación semántica permitirá detectar a cada uno de los píxeles que corresponden a los edificios que se encuentren en la imagen y los agrupará dentro de la misma clase (figura 11).

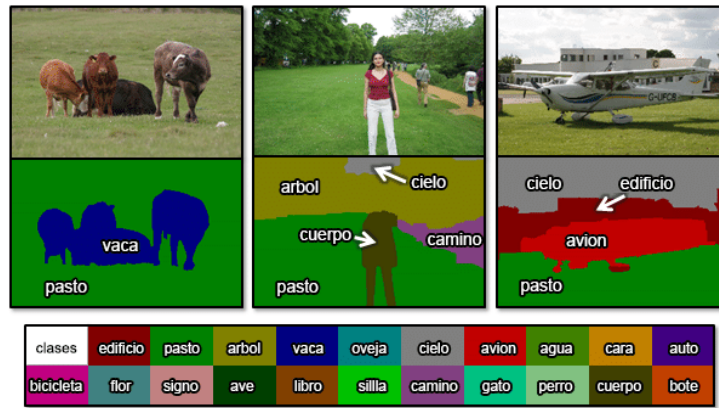


FIGURA 11. SEGMENTACIÓN SEMÁNTICA APLICADA A LA DETECCIÓN DE OBJETOS.

La segmentación semántica tiene como objetivo asignar a cada uno de los pixeles una clase diferente como pueden ser edificio, pasto, árbol, vaca, oveja, cielo, avión, agua, cara, auto, bicicleta, flor, signo, ave, libro, silla, camino, gato, perro, cuerpo y bote como se muestra en la figura. Fuente: https://www.researchgate.net/figure/Figura-24-Ejemplo-de-Segmentacion-Semantica-31_fig4_319766387.

La segmentación semántica ha encontrado su uso en el diagnóstico de algunas enfermedades o tumores en biomedicina (figura 12), separación de células, visualización en carros autónomos, separación de fondo en imágenes o videos, detección del ambiente externo de robots y segmentación de expresiones para separar objetos que pertenecen a la misma clase, e incluso poder separar clases dentro de clases como muestra en la figura 13, donde se puede identificar primero a la clase “personas” y después separar aquellas que contienen un elemento en común como utilizar el abrigo del mismo color [57].

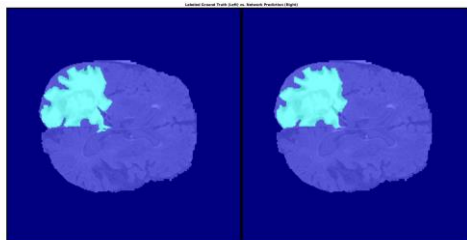


FIGURA 12. SEGMENTACIÓN SEMÁNTICA APLICADA EN BIOMEDICINA. Identificación de un tumor cerebral con segmentación semántica. Fuente [58].

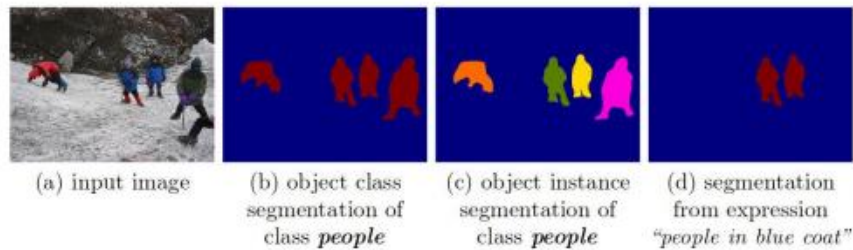


FIGURA 13. SEGMENTACIÓN SEMÁNTICA SEPARANDO OBJETOS DE MISMA CLASE.

Segmentación que permite identificar objetos de la misma clase, eliminando los de color diferente. Fuente [59].

La idea principal de la segmentación semántica está basada en codificar la información contenida en la imagen para posteriormente decodificarla, respeta las fronteras que existen entre los objetos que pertenecen a una imagen. Así, más que un proceso de segmentación o clasificación es un proceso de extracción de características de los objetos que se encuentran dentro de la imagen, y los identifica como pertenecientes a diferentes clases. El uso de características locales permite definir dos cosas: El punto de interés y la descripción de la región de interés.

El proceso de segmentación semántica permite hacer la asociación entre cada píxel que se encuentra contenido dentro de la imagen, con una clase que ha sido etiquetada y previamente definida (como puede ser una flor, una banda de proteína o el cielo). El empleo de la segmentación semántica sobre los píxeles permite conseguir múltiples vectores de características que, facilitan la interpretación de la información conceptual de las relaciones presentes entre los objetos de la imagen, por lo que las características diferentes de cada objeto no necesitan ser extraídas de la imagen y se les puede asociar a un término semántico o una clase bien definida [60, 59].

El proceso de codificación en la segmentación semántica inicia con el empleo de convoluciones a la imagen para poder recuperar las características más importantes que contiene la imagen. La red neuronal convolucional o CNN es

una red profunda con una arquitectura de retroalimentación que permite comparar los datos en diferentes capas permitiendo el aprendizaje de características abstractas para que por medio de filtros se pueda identificarlas posteriormente y de manera eficiente.

Posterior al proceso de convolución, la red se prepara para aplicar procesos no lineales empleando una unidad lineal rectificadora o función ReLU consiguiendo la eliminación de los valores negativos y manteniendo solamente los valores positivos.

Para reducir el tamaño de la matriz de píxeles se emplea el proceso de “*pooling*” consiguiendo disminuir el número de procesamientos realizados por computadoras para tener un proceso eficiente al momento de tratar los datos y facilitar el empleo de filtros para reconocer las características principales de la imagen durante la aplicación de las redes neuronales convolucionales.

Este proceso se repite tantas veces sea necesario para que las neuronas involucradas en la CNN aplicada vayan entrenándose, aprendiendo las características de la imagen para ir generando los diferentes tipos de clases que pueden extraerse de la misma imagen como se muestra en la figura 14 [61].

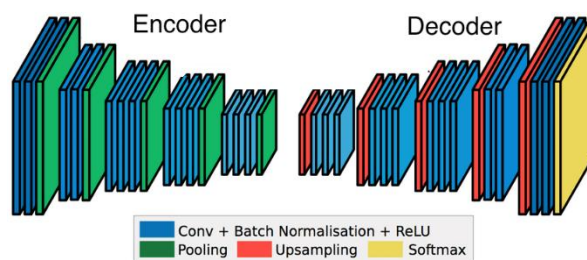


FIGURA 14. REPRESENTACIÓN DE UNA CNN APLICADA A LA SEGMENTACIÓN SEMÁNTICA.

Fuente [61].

Cada paso realizado durante la codificación de la información deberá ir relacionado con su respectiva decodificación para extraer posteriormente los

valores que pertenecen a la imagen que se encuentre analizando la red empleando la arquitectura *U-net*, este proceso permitirá generar las clases correspondientes que han sido previamente definidas. Este tipo de arquitectura ya ha sido empleada para la segmentación de imágenes médicas consiguiendo mejor precisión que otras arquitecturas (figura 15).

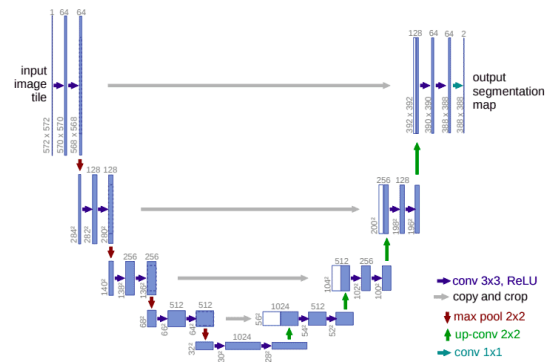


FIGURA 15. ARQUITECTURA U-NET EMPLEADA PARA EL ANÁLISIS DE IMÁGENES MÉDICAS.

Fuente: [62].

Finalmente, la decodificación genera la salida volviendo a ensamblar los píxeles para conseguir la imagen segmentada y dividida en las clases requeridas para su análisis final [62, 63, 64, 65].

Capítulo 3. Solución Propuesta

La metodología de investigación desarrollada para este estudio se encuentra dividida en dos partes.

La primera parte consiste en la expresión de la proteína GPN de forma recombinante y se utilizó para aplicar un método de purificación que permitió conseguir la proteína lo más pura posible, homogénea y con menor polidispersión. La proteína pura se centrifugó a diferentes concentraciones para mezclarla con extractos celulares y emular distintas muestras que representaron a pacientes con diferentes estadios de cáncer. Los prototipos generados se utilizaron para desarrollar los geles SDS-PAGE que incluyeron muestras sin la proteína y con diferentes cantidades de sobreexpresión de la GPN y se emplearon para obtener las imágenes deseadas para el estudio.

La segunda parte explica las técnicas de análisis de imágenes y de reconocimiento de patrones para eliminar el ruido de fondo y conseguir la segmentación de las bandas de proteínas presentes en el gel SDS-PAGE. Esta metodología permitió encontrar de forma automática la cantidad de muestras presentes en el gel, la posición de la proteína de interés en base a su peso molecular y la sobreexpresión de la proteína.

Las dos partes se detallan a continuación.

3.1. Emulación de geles de pacientes con diferente sobre expresión de proteína

Para la obtención de la proteína de forma soluble, pura, homogénea y monodispersa; que fue utilizada para el desarrollo de los geles SDS-PAGE,

que permitieron generar la base de imágenes de la expresión y no expresión de la proteína GPN se aplicaron los siguientes pasos (figura 16).

- a) Se produjo el vector de expresión del gen de la proteína GPN para expresarlo en bacterias.
- b) Se insertó el vector en la bacteria *Escherichia coli* y se expresó la proteína recombinante.
- c) Se obtuvieron geles de poliacrilamida de la expresión y purificación de la proteína GPN que fueron revelados en azul de Coomasie.
- d) Se desarrollaron diferentes buffers para mantener la proteína GPN en óptimas condiciones.
- e) Se consiguió el buffer óptimo que mantuvo la proteína pura, homogénea y con un grado de polidispersión del 20% cotejado por un estudio de dispersión dinámica de luz (DLS).
- f) Se agregó la proteína purificada a extractos celulares utilizando distintas concentraciones para conseguir muestras sin expresión y a diferentes niveles de expresión de la proteína GPN para emular los estadios de cáncer relacionado con la sobreexpresión de la proteína en estudio.
- g) Se creó la base de imágenes de los geles SDS-PAGE divididos en no expresión y sobreexpresión de la proteína GPN que fueron analizados por técnicas de análisis de imágenes y segmentación semántica.

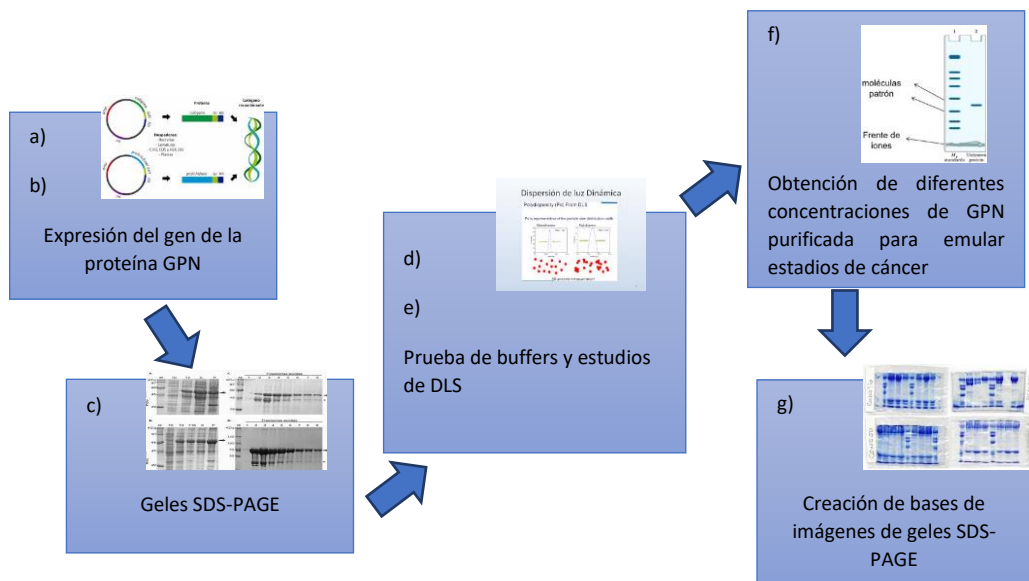


FIGURA 16. DIAGRAMA DE FLUJO DE LA PRIMERA PARTE DE LA METODOLOGÍA.
Pasos requeridos para generar la base de datos. Fuente: Elaboración propia.

Los detalles de la metodología aplicada se explican en las siguientes secciones.

3.1.1. Construcción del vector de expresión para la GPN

Los genes que codifican la proteína humana GPN fueron obtenidos de la librería de *cDNA human Open Biosystems, US* y se clonaron junto con un grupo de seis histidinas (hexahistidinas) dentro del vector pET28-DL3 y se insertó dentro de la bacteria *Escherichia coli* siguiendo la metodología propuesta por González [66].

3.1.2. Preparación de la columna His-bind

Para la preparación de la columna His-bind se usaron 100 μ l de resina del kit His-Bind de Novagen mezclado con 25 μ l de sulfato de níquel y 175 μ l del buffer 100 mM Tris-HCl con 100 mM de KCl manteniendo un pH de 8.2. Esta preparación fue colocada dentro de un tubo Eppendorf que se mantuvo en rotación durante un tiempo de 15 minutos para equilibrar la muestra. Pasado el tiempo establecido, el tubo Eppendorf se lavó empleando 10 ml del mismo buffer y se utilizó como columna para retener la proteína recombinante.

3.1.3. Expresión de la proteína GPN y crecimiento a diferentes temperaturas

La cepa bacteriana *Escherichia coli* (DL3) transformada con el vector de expresión del gen de la proteína GPN fue mantenida a 37°C para su crecimiento dentro de un frasco de 1 litro con 600 ml de medio de cultivo LB a 100 μ M/ml de Kanamicina. Una vez alcanzada la fase estacionaria se indujo la expresión de la proteína recombinante agregando *Isopropyl β -D-1-thiogalactopyranoside* (IPTG) a una concentración de 200 μ M. Se tomaron dos muestras de cultivo dejándolas en crecimiento durante 24 horas, la primera se mantuvo a una temperatura de 16°C, denominándola con la clave MG16C y la segunda a 10°C, clave MG10C. Transcurriendo las 24 horas los cultivos celulares MG16C y MG10C fueron centrifugados en tubos Eppendorf a una temperatura de 4°C a 13,000 rpm por un tiempo de 15 minutos y los pellets fueron resuspendidos en 1 ml de sus respectivos buffers a pH 8.2. Las soluciones fueron sonicadas tres veces con pulsos de 30 segundos por 30 segundos de reposo en hielo con sal y acetona. Posteriormente fueron centrifugadas a una temperatura de 4°C a 13,000 rpm durante 15 minutos.

Los sobrenadantes se colocaron en los 100 μ l de resina de la columna His-Bind preparada previamente y equilibrada con níquel, las muestras se mezclaron y se dejaron rotar durante 20 minutos. Terminando la rotación las columnas se lavaron con 10 ml de sus respectivos buffers a pH 8.2 para finalmente eluir las con el mismo buffer agregando 300 mM de imidazole para separar la proteína recombinante de las columnas. Finalmente, las muestras se pasaron por una columna de desalado 6K MWCO de 10 ml *Thermo Scientific* para eliminar las sales y se separaron empleando electroforesis SDS-PAGE al 10% para revelarse mediante la técnica de tinción de azul de Coomassie (ver proceso en el diagrama de la figura 17).

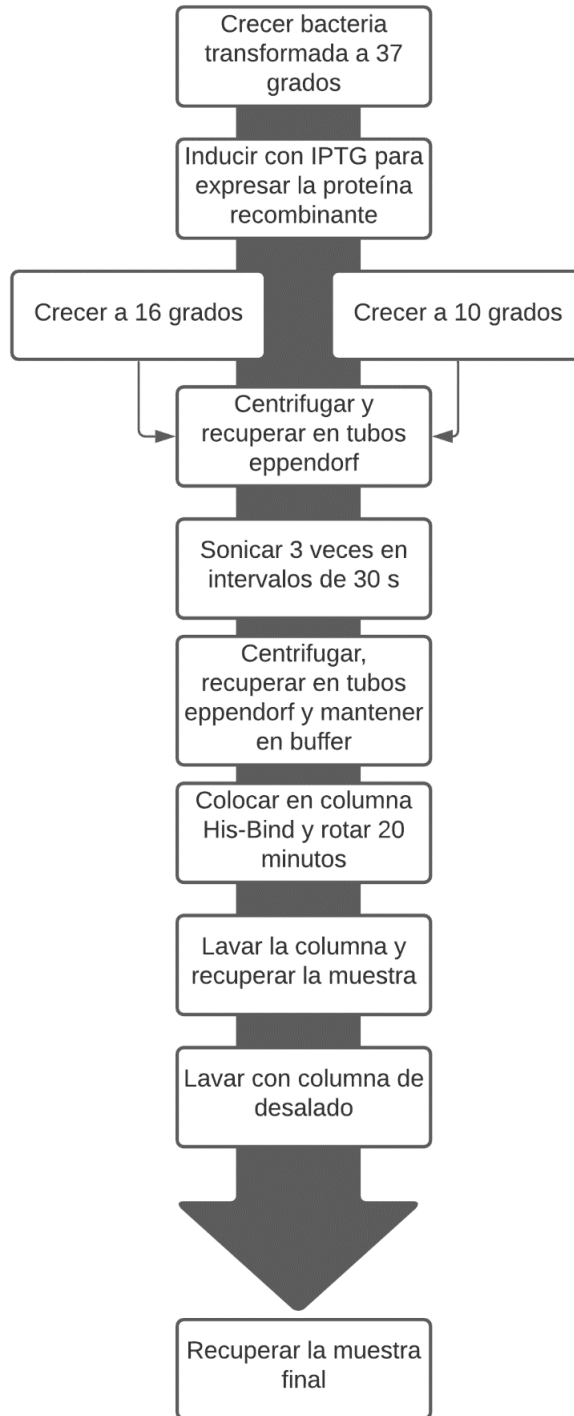


FIGURA 17. DIAGRAMA DE FLUJO DEL PROCEDIMIENTO PARA CONSEGUIR LA MUESTRA FINAL.
Fuente: Elaboración propia.

3.1.4. Preparación a diferentes concentraciones de buffer con NaCl

Se prepararon cinco buffers con 50, 100, 500, 1000 y 2000 mM de NaCl agregados a 100 mM Tris-HCl a pH 8.2. Estos se utilizaron como buffers para realizar la expresión y purificación de la proteína GPN con la muestra MG10C siguiendo el procedimiento descrito anteriormente y resumido en el diagrama de la figura 17. Cada uno de los buffers fueron comparados para encontrar el óptimo en conseguir la pureza de la proteína recombinante.

Mediante el análisis de Bradford se calculó la concentración obtenida de las proteínas del extracto total en cada una de las muestras tratadas con los buffers preparados.

3.1.5. Preparación a diferentes concentraciones de buffer con TRIS

Se prepararon cuatro buffers modificando la concentración de Tris a 50, 100, 150 y 200 mM con 100 mM NaCl manteniendo un pH de 8.2 y se emplearon en el proceso la purificación de la proteína GPN con la muestra MG10C. para localizar el óptimo.

3.1.6. Preparación a diferentes concentraciones de buffer con EDTA

Se prepararon tres muestras diferentes empleando las concentraciones de 100, 200 y 400 μ M de EDTA y se agregaron al buffer 100 mM Tris-HCl, 100 mM de NaCl a un pH de 8.2 y se emplearon para la purificación de la proteína GPN utilizando la muestra MG10C separando el óptimo.

3.1.7. Preparación a diferentes concentraciones de buffer con DTT

Se prepararon 6 soluciones buffer a base de 100 mM Tris-HCl, 100 mM NaCl a un pH 8.2 con 10, 20, 40, 60, 80 y 100 μ M de DTT en cada uno de los buffers y se utilizaron para el proceso de la purificación de la proteína GPN con la muestra MG10C eligiendo el óptimo.

3.1.8. Purificación con aminoácidos a diferentes concentraciones

Se repitió el procedimiento de purificación de la proteína GPN utilizando la muestra MG10C con las diez soluciones buffer tomando como base 100 mM Tris-HCl, 100 NaCl a pH 8.2 y las siguientes concentraciones: 1) 25 mM Glutamato de Sodio, 2) 50 mM Glutamato de Sodio, 3) 100 mM Glutamato de Sodio, 4) 200 mM Glutamato de Sodio, 5) 300 mM Glutamato de Sodio, 6) 50 mM Arginina, 7) 50 mM Lisina, 8) 100 mM Lisina, 9) 200 mM Lisina, 10) 100 mM Lisina, 100 mM Glutamato de Sodio.

Se compararon las muestras obtenidas y se eligió la que mantuvo la proteína con mayor pureza.

3.1.9. Purificación con Buffer Final

Se realizó el proceso de purificación de la proteína GPN con la muestra MG10C empleando el buffer final obtenido al combinar los amortiguadores óptimos conseguidos mediante la realización de los experimentos descritos anteriormente (secciones 3.1.3 hasta la 3.1.8), manteniendo las siguientes

soluciones: 100 mM Tris-HCl 100, 5% glicerol, 0.2% Tritón X-100, 100 mM NaCl, 100 mM lisina, 100 mM Glutamato de Sodio, 400 μ M EDTA, 100 μ M DTT empleando el sistema IMAC descrito anteriormente.

3.1.10. Cuantificación de la proteína GPN

Para calcular la concentración de la proteína, se cuantificó empleando el kit de Bio-Rad (Bio-Rad Bradford *protein assay*) siguiendo las especificaciones del proveedor. Las mediciones se realizaron utilizando una absorbancia de 595 nm para construir la curva de calibración y con ella estimar el peso molecular de la proteína purificada.

3.1.11. Análisis de dispersión de luz dinámica

Los experimentos de dispersión dinámica fueron realizados en un instrumento Malvern Nano S (Malvern, Ltd) equipado con tecnología láser NIBS (*Non-Invasive Back Scattering*) a una longitud de onda de 663 nm con un controlador de temperatura Peltier. La solución de proteínas fue filtrada a través de un filtro de jeringa Anotop® con un tamaño de poro de 0.02 μ m (Whatman, GE) antes de realizar las mediciones. El radio hidrodinámico (Rh) fue calculado empleando el software Zeta Sizer proporcionado por el mismo equipo.

3.1.12 Emulación de muestras con diferentes niveles de sobre expresión de la proteína GPN para representar distintos estadios de cáncer de mama relacionados con la sobre expresión de la proteína

Para la creación de muestras con diferentes concentraciones de proteína GPN que puedan emular la proteína sobre expresada tal y como ocurre con la proteína in vivo durante la afección de cáncer de mama CDI y CLI de tipo Her2+ [27] se siguió la metodología de purificación de cromatografía de afinidad por metales inmovilizados desarrollada por [40] para la obtención de proteína GPN de forma pura, homogénea y monodispersa y se concentró como lo explica [67, 40].

Se tomó extracto celular a partir de la bacteria *Escherichia coli* (*E. coli*) y se mezcló con la proteína purificada (ver diagrama de la figura 17) para obtener muestras con las siguientes concentraciones: 0, 2, 9.5, 14, 14.5, 18, 18.5, 26, 27 y 30 mg/ml respectivamente.

También se preparó una dilución de la proteína Albúmina de Suero Bovino (BSA) obtenida del kit Bio-Rad *Protein assay* para conseguir tres muestras con las cantidades de 2 mg/ml, 1 mg/ml, y 0.5 mg/ml.

Todas las muestras fueron separadas empleando electroforesis SDS-PAGE (*Sodium Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis*) al 10% y se reveló el gel obtenido empleando la técnica de azul de Coomasie.

El resumen de la técnica completa se ilustra en el diagrama de la figura 18, que presenta la obtención de muestras con proteína a diferentes concentraciones, su electroforesis y el revelado con azul de Coomasie.

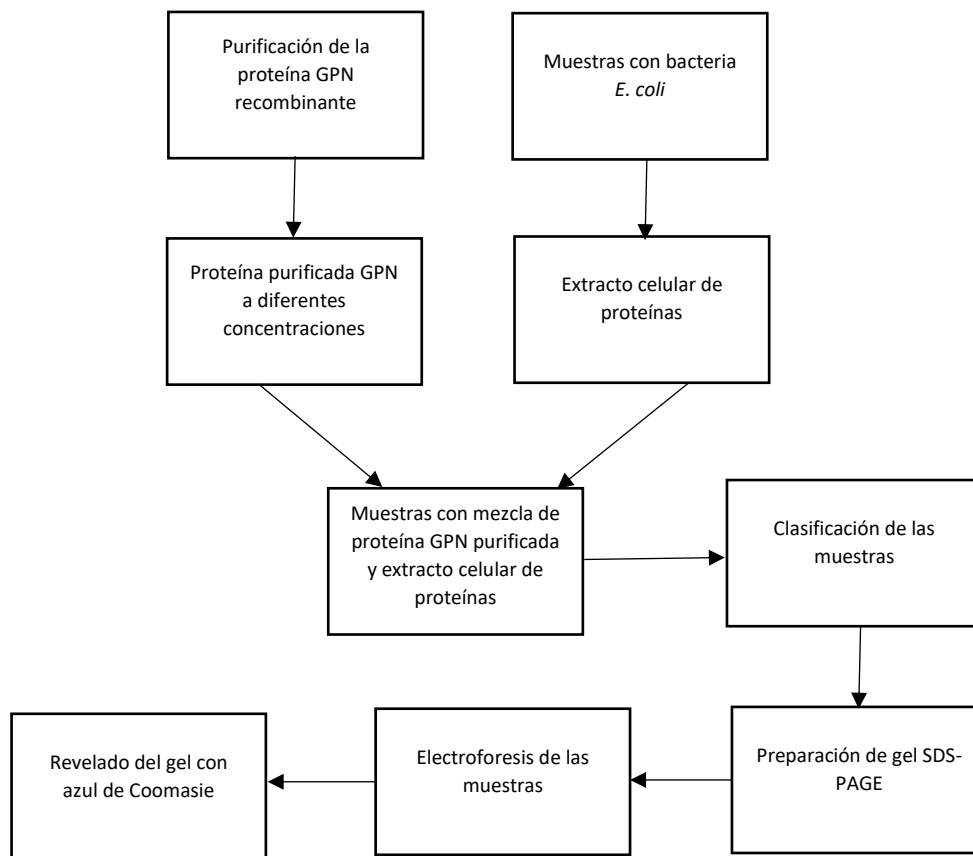


FIGURA 18. DIAGRAMA PARA OBTENER MUESTRAS EN GELES SDS-PAGE.
Fuente: Elaboración propia.

3.2. Desarrollo de algoritmos de análisis de imágenes y segmentación semántica

Para el desarrollo de los algoritmos que detectan la cantidad de muestras en el gel, el número de bandas, su peso molecular y sobreexpresión de la proteína de interés se utilizaron los siguientes pasos (figura 19).

- a) Se clasificó la base de imágenes de geles para dividir las imágenes de los geles obtenidos en dos grupos: 1) GPN no expresada, 2) GPN expresada a diferentes niveles.

- b) Se ajustaron los colores, la nitidez y el brillo de las imágenes mediante el preprocesamiento de los datos. Se buscaron técnicas de eliminación de ruido de los cambios de coloración provocados por la tinción del gel de poliacrilamida para conseguir su homogeneidad y tener la misma calidad en el color.
- c) Se binarizaron las imágenes para cambiarlas a blancos y negros para poder realizar los estudios de clasificación y facilitar la detección de las bandas de proteínas.
- d) Se emplearon técnicas de erosión, dilatación y umbralización para mejorar el contorno de las bandas que representan a las proteínas dentro de un gel de poliacrilamida SDS-PAGE.
- e) Se eliminó el ruido de fondo, se emplearon filtros, transformadas y correlaciones hasta minimizar o desaparecer el ruido de fondo y eliminar falsos positivos.
- f) Se emplearon algoritmos de segmentación y transformadas que permitieron la identificación de las líneas que delimitan los diferentes experimentos o muestras que contiene el gel.
- g) Se emplearon algoritmos de segmentación y transformadas que permitieron la identificación de la banda que corresponden a la proteína de interés biomédico.
- h) Se extrajeron características que permitieron la identificación de la posición de las bandas de proteínas empleando segmentación semántica.
- i) Se compararon métodos tradicionales de análisis con los obtenidos por medio de la segmentación semántica.

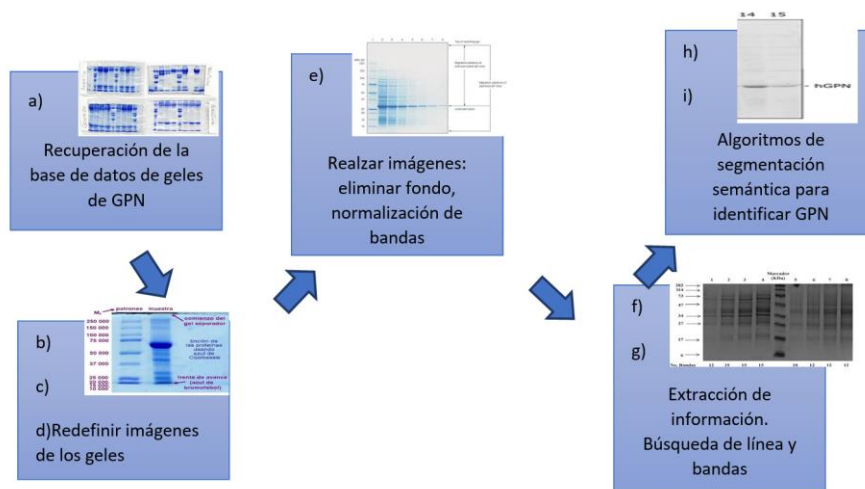


FIGURA 19. DIAGRAMA DE FLUJO DE LA SEGUNDA PARTE DE LA METODOLOGÍA.

Pasos requeridos para encontrar el número de muestras, la banda de proteína GPN, su peso molecular y la sobreexpresión de la proteína. Fuente: Elaboración propia.

Los detalles de la metodología desarrollada se explican a continuación.

3.2.1. Adquisición de la imagen

Las imágenes de los geles fueron obtenidas empleando un sistema fotodocumentador Gel Doc XR+ basado en cámaras de alta resolución empleando el software *image Lab* para capturar las imágenes siguiendo las especificaciones del proveedor Bio-Rad, la resolución de las imágenes fueron de 4 megapíxeles, tamaño de píxel (H x V) 4.65 x 4.65 μM y una densidad de píxeles en nivel de gris de 4,096 (<https://www.bio-rad.com/es-mx/product/gel-doc-xr-gel-documentation-system?ID=O494WJE8Z>).

3.2.2. Preprocesamiento y extracción de características empleando una nueva metodología denominada perfil de imagen basada en segmentación de imágenes binarias (PIBSIB)

Las imágenes de los geles obtenidas con el fotodocumentador fueron pretratadas tal y como se especifica en [67], para ello, primero fueron

redimensionadas a un tamaño de 600 x 400 píxeles (px). Las imágenes fueron ecualizadas solamente en el caso de que las muestras presentaran un exceso de proteínas. Posteriormente, todas las imágenes se binarizaron y se dilataron (ver figura 20). Se invirtieron colores y se erosionaron.

Se realizó primero el análisis correspondiente a las líneas (el número de muestras dentro del gel) y después a las bandas (la cantidad de proteínas por muestra). Si el análisis se realizó sobre las líneas la máscara binaria empleada tuvo una dimensión de 1x400 píxeles, en cambio, para el estudio de las bandas la máscara binaria tuvo una dimensión de 1x50 píxeles.

La máscara binaria se aplicó a la imagen analizada en la región correspondiente (1x400 px ó 1x50 px), se calculó el histograma de la región contenida dentro de la máscara binaria y se evaluaron diferentes técnicas de binarización (entre las que se incluyen Niblack, Sauvola, *thresholding* y Otsu sin que ocurriera variación en los resultados); de los valores obtenidos solo se utilizó el que corresponde al píxel blanco (posición 255) y se almacenó en un arreglo. La máscara avanzó píxel por píxel hasta terminar de evaluar los 600 píxeles que corresponden a las columnas de la imagen analizada. Se graficó el arreglo final obtenido con los diferentes valores de intensidad del color blanco para generar un nuevo perfil donde cada uno de los múltiples mínimos correspondieron a las separaciones de las distintas muestras que contiene el gel, siempre y cuando el análisis realizado fuera sobre las columnas. En caso de que el análisis correspondiera a las bandas de una sola muestra los múltiples máximos coincidieron con la posición de las proteínas presentes (por lo que se promediaron para conseguir un solo máximo). Ver diagrama en la figura 20 y el algoritmo de la metodología en la figura 21.

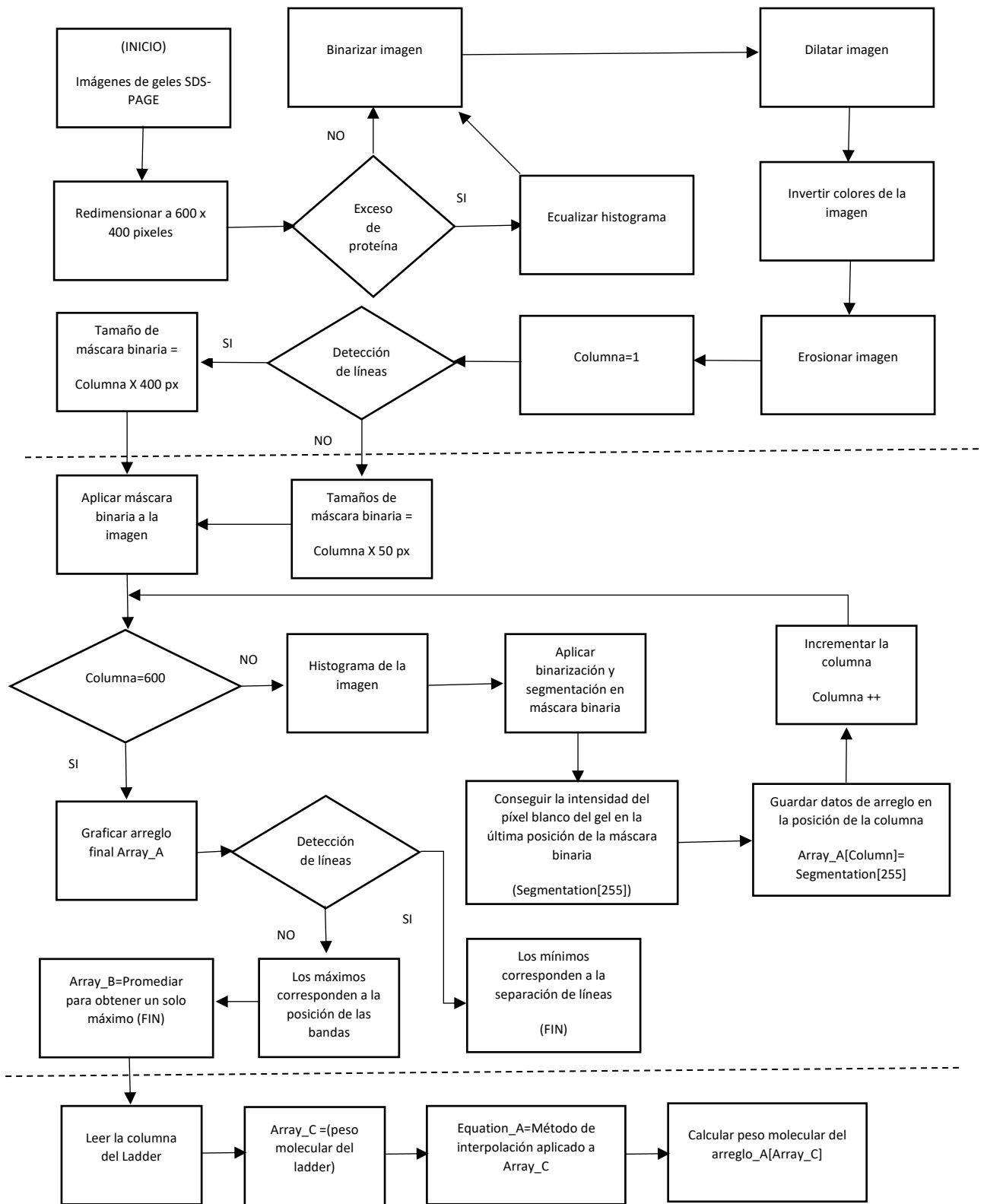


FIGURA 20. ESQUEMA GENERAL. FUENTE: ELABORACIÓN PROPIA.

Algoritmo desarrollado para el análisis de imágenes de los geles de poliacrilamida.

Algoritmo para la detección de líneas y bandas

```
1: Redimensionar la imagen a to 600 x 400 px
2: if Exceso_de_proteína:
3:     Ecualizar histograma
4: end if
5: Binarizar imagen
6: Dilatar imagen
7: Invertir colores de imagen
8: Erosión de imagen
9: Columna=1
10: If Detección_Línea:
11:     Tamaño_Máscara_Binaria =Columna X 400 px
12: else:
13:     Tamaño_Máscara_Binaria =Columna X 50 px
14: end else
15: end if
16: Aplicar Máscara Binaria
17: Array[]
18: while Columna <=600:
19:     Histograma_de_Imagen
20:     Aplicar_Binarizacion_Segmentación_En_
    La_Región_de_la_Máscara_Binaria
21:     Array[Columna]=Intensidad_Pixel_Blanco_[255]
22:     Columna++
23: end while
24: Graficar Array
```

```

25:  if Detección_Línea:
26:     Múltiples_Mínimos_corresponden_a_Separación_Líneas
27:  else:
28:     Múltiples_Máximos_corresponden_a_Separación_Bandas
29:     Promediar_Múltiples_Máximos_Para_Conseguir_El_Máximo
30:  end else
31:  end if

```

FIGURA 21. . ALGORITMO GENERAL.

Metodología que permite encontrar de forma automática el número de muestras presentes en el gel de poliacrilamida y las bandas que corresponden a las proteínas de una muestra. Fuente: Elaboración propia.

3.2.3. Extracción de características empleando segmentación semántica

La segmentación semántica se realizó con el software Matlab R2022a© empleando aprendizaje profundo para una región rectangular utilizando 100 imágenes como prueba, 100 como etiqueta, 200 para entrenamiento y 400 como etiquetas de entrenamiento. Todas las imágenes fueron empleadas para realizar el entrenamiento de la red y detectar las clases correspondientes a las bandas de proteínas y al fondo o ruido.

Las imágenes de los geles empleados incluyeron geles con diferente tonalidad de color, geles rotos o mal preparados.

La red utilizada incluyó las siguientes capas con los siguientes elementos proporcionados por Matlab R2022a© tal como se muestra en la figura 22:

1. 2-D *convolution*.
2. ReLU.

3. 2-D *max pooling*.
4. 2-D *convolution*.
5. ReLU.
6. 2-D *transposed convolution*.
7. 2-D *convolution*.
8. Softmax.
9. Pixel *classification layer*.

Se utilizó el mismo software para encontrar la exactitud del análisis realizado por la segmentación semántica.

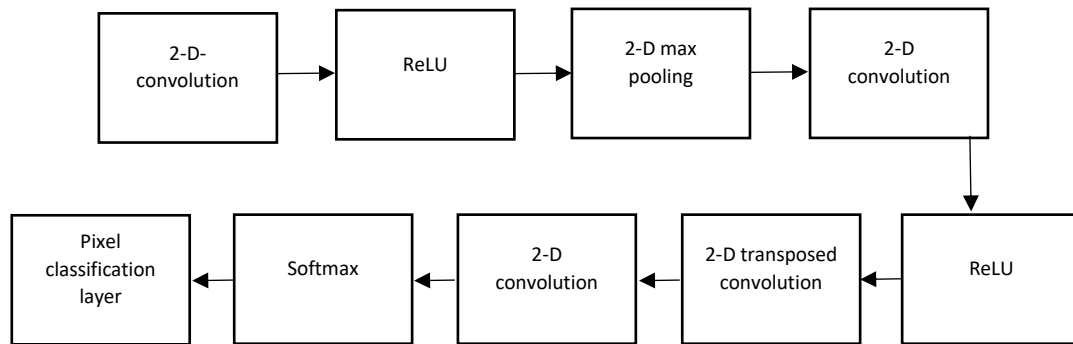


FIGURA 22. ESQUEMA GENERAL DE RED NEURONAL.

Red utilizada para la realización de la segmentación semántica empleando Matlab R2022a©.
Fuente: Elaboración propia.

Capítulo 4. Resultados obtenidos

Después conseguir la proteína GPN en forma pura, se desarrollaron los geles que emularon los diferentes estadios de la enfermedad de cáncer relacionada con la sobreexpresión de la proteína de interés. Se construyó la base de imágenes dividida en muestras con y sin proteína GPN para realizar el análisis de imágenes que permitió identificar de forma semi-automática la cantidad de muestras presentes. Se logró reconocer las bandas por su peso molecular y se comparó la cantidad expresada de la proteína. Los resultados obtenidos se muestran a continuación.

4.1. Construcción del vector de expresión para la GPN y prueba de expresión

Para verificar que la construcción del vector pudiera expresar la proteína GPN dentro de la bacteria *Escherichia coli* (*E. coli*) se tomó 1 ml del cultivo de bacteria antes de inducir su crecimiento con IPTG (figura 23, línea 3) y 1 ml después de la inducción (figura 23, línea 2) y se separó empleando electroforesis SDS-PAGE al 10% para ser revelado empleando la técnica de tinción de azul de Coomasie. El gel de la figura 23 en la línea 2 indica que se obtuvo buena expresión de la proteína GPN al visualizarse una banda con mayor grosor comparada con las proteínas del extracto total. Además, se puede verificar que la posición de dicha banda en el control positivo (línea 2) tiene un peso molecular aproximado de 34 Kilo Daltones o KDa (comprendido entre los 25 y 37 KDa que muestra el control o marcador de peso molecular o Ladder en la línea 1), y que es el valor predicho para su peso molecular. La banda correspondiente a la GPN no se muestra en la línea 3 o control negativo donde no se realizó su expresión. Estos resultados corroboran que el gen de la GPN (vector de expresión) que se insertó en la bacteria realmente contiene la proteína de interés.

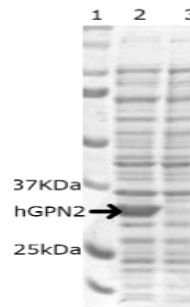


FIGURA 23 PRUEBA DE EXPRESIÓN DE PROTEÍNA GPN.

Análisis SDS-PAGE de la proteína recombinante. Línea 1, marcador de peso molecular de proteínas. Línea 2, proteína GPN recombinante expresada en *E. coli*. Línea 3, extracto total de proteína de *E. coli* sin la expresión de la proteína GPN. Fuente: Elaboración propia.

4.2. Preparación de la columna His-bind

Para la creación de la columna His-bind se prepararon 28 tubos Eppendorf con 100 μ l de resina His-bind y 25 μ l de sulfato de níquel y se almacenaron en refrigeración a 4°C para ser utilizados como columnas cromatográficas de afinidad a metales inmovilizados IMAC durante el proceso de purificación y crecimiento de la proteína GPN a diferentes temperaturas con cada uno de los buffers mencionados en la metodología.

4.3. Expresión de la proteína GPN y crecimiento a diferentes temperaturas

Después de la inducción con IPTG se realizaron dos crecimientos manteniendo diferentes temperaturas (figura 17). La muestra con la bacteria se colocó en un tubo Eppendorf y se centrifugó durante 15 minutos a 4°C, se dividió el sobrenadante del pellet y se realizó la separación mediante

electroforesis SDS-PAGE al 10% para revelarse por medio de tinción con azul de Coomasie. La figura 24A muestra que en el crecimiento a 16°C (MG16C) la proteína recombinante se mantuvo en el pellet por lo que no se encuentra en la fracción soluble. Al realizar el crecimiento disminuyendo la temperatura a 10°C (MG10C) el pellet muestra una disminución en la cantidad de la proteína GPN debido a que aparece en la parte soluble contenida en el sobrenadante (ver figura 24B, línea 2), por lo que el análisis del lisado celular y del medio del cultivo mostró una máxima expresión de la proteína recombinante a las 24 horas de crecimiento después de la inducción manteniendo la temperatura a 10°C, lo que hace que la muestra MG10C permita conseguir una mayor cantidad de la proteína recombinante de forma soluble.

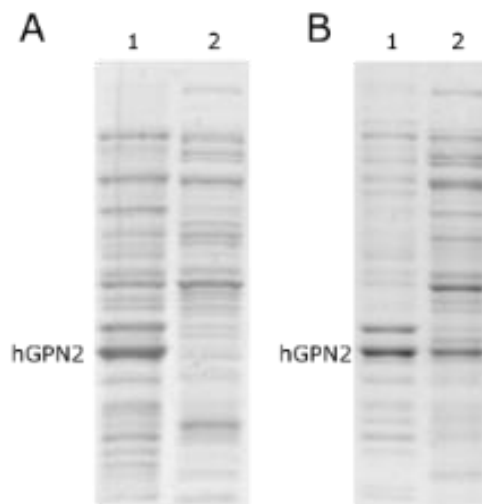


FIGURA 24. PRUEBA DE EXPRESIÓN DE LA PROTEÍNA GPN A DIFERENTES TEMPERATURAS.

A) Línea 1 pellet, línea 2 sobrenadante sin GPN expresada con un crecimiento a 16 grados centígrados. B) Línea 1 pellet, línea 2 sobrenadante con GPN expresada en la fracción soluble con un crecimiento a 10 grados centígrados. Fuente: Elaboración propia.

4.4. Análisis con diferentes concentraciones de NaCl

Los cambios de concentración de sal (NaCl) en un medio acuoso tienen el efecto en las proteínas de precipitarlas o incrementar su solubilidad. Para encontrar la cantidad de concentración salina que favorece una mayor concentración de proteína GPN soluble se llevó a cabo la medición de diferentes concentraciones de cloruro de sodio (NaCl) en la preparación de un buffer básico, que no tuviera ningún componente más que 100 mM de Tris-HCl para mantener el pH a 8.2, ya que es el requerido para las funciones biológicas de la proteína en estudio. La muestra MG10C presentó los valores que se grafican en la figura 25.

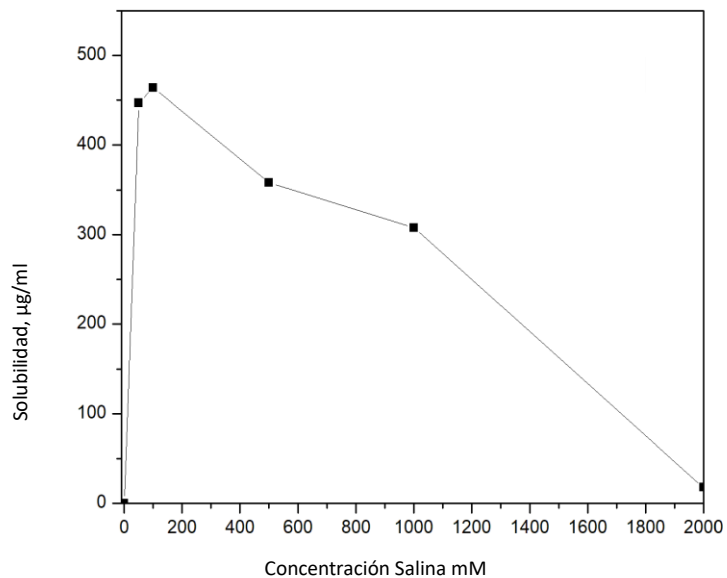


FIGURA 25. ANÁLISIS CON DIFERENTES CONCENTRACIONES DE NaCl.

Gráfica que presenta cinco muestras de concentración salina a 50, 100, 500, 1000 y 2000 mM para encontrar la máxima cantidad soluble de proteína GPN. Los valores muestran que después de alcanzar los 100 mM de NaCl la proteína precipita dejando de ser parte de la fracción soluble. Fuente: Elaboración propia.

Para encontrar la concentración de proteína soluble se realizó el análisis de Bradford a cada una de las diferentes concentraciones salinas obteniendo los datos de solubilidad mostrados en la tabla 2.

Los valores de solubilidad obtenidos indicaron que la mayor concentración de proteína GPN se consiguió al emplear el buffer 100 mM Tris-HCl, 100 mM NaCl. Al utilizar concentraciones superiores la solubilidad fue disminuyendo, mostrando que a concentraciones mayores de 2000 mM de NaCl el contenido de la proteína recombinante y del extracto total precipita por lo que deja de mantenerse en la fracción soluble.

TABLA 2. SOLUBILIDAD OBTENIDA POR CONCENTRACIÓN SALINA.

Diferentes concentraciones de cloruro de sodio mantenidas en cinco buffers y su solubilidad obtenida en cada una de las muestras medidas con el análisis de Bradford.

Concentración salina (Molar)	Solubilidad µg/ml
0	0
50	447
100	464
500	358
1000	308
2000	18

4.5. Análisis con diferentes concentraciones de TRIS

Se midió el efecto de las diferentes concentraciones de Tris manteniendo constante el NaCl a 100 mM adicionado al buffer a pH 8.2. En la comparación

realizada con Tris se puede observar que en el proceso de la purificación de la proteína GPN de la muestra MG10C, las concentraciones menores de 100 mM (figura 26, línea 1) conservaron altos valores de contaminantes representados por bandas con mayor grosor comparadas con las bandas obtenidas en los experimentos subsecuentes, y que al utilizar valores mayores a 100 mM (figura 26, líneas 2-4) ya no existieron variaciones en la cantidad de proteínas manteniendo los mismos valores en el extracto celular analizado.

Los resultados obtenidos indicaron que el Tris-HCl no tiene efecto significativo en la eliminación de impurezas y que una concentración de 100 mM es la cantidad mínima de Tris-HCl requerida para mantener el pH y disminuir levemente la cantidad de contaminantes sin presentar una reducción en el tamaño de la banda que corresponde a la GPN.

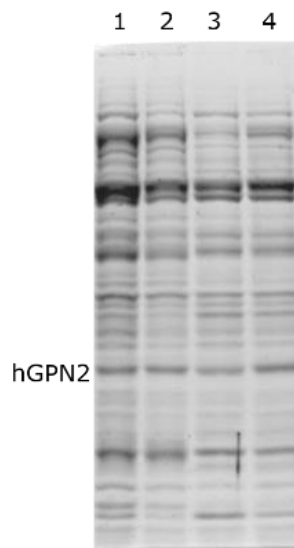


FIGURA 26. PURIFICACIÓN DE LA PROTEÍNA RECOMBINANTE GPN EN DIFERENTES BUFFERS: Línea 1 50 mM Tris-HCl, 100 mM NaCl. Línea 2 100 mM Tris-HCl, 100 mM NaCl. Línea 3 150 mM Tris-HCl, 100 mM NaCl. Línea 4 200 mM Tris-HCl, 100 mM NaCl (línea 4). Fuente: Elaboración propia.

4.6. Análisis con diferentes concentraciones de EDTA

Se prepararon tres muestras de MG10C para purificar la proteína recombinante GPN. La figura 27 muestra como el empleo de 400 μM de EDTA reduce la cantidad de contaminantes durante la purificación (figura 27, línea 3), valores menores (100 y 200 μM) no consiguen la disminución de la cantidad de contaminantes ya que aparecen bandas de proteínas no deseadas de mayor grosor (figura 27, líneas 1-2). Estos resultados sugieren que el empleo de 400 μM de EDTA favorece la disminución de contaminantes, aunque sin conseguir eliminarlos todos.

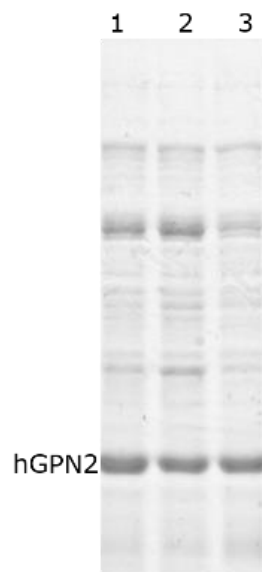


FIGURA 27. ANÁLISIS CON DIFERENTES CONCENTRACIONES DE EDTA.

Expresión de la proteína GPN empleando el buffer 100 mM Tris-HCl, 100 mM NaCl, manteniendo pH a 8.2 y concentraciones de EDTA de 100 μM (línea 1). 200 μM (línea 2). 400 μM (línea 3). Fuente: Elaboración propia.

4.7. Análisis con diferentes concentraciones de DTT

Al utilizar DTT junto con la muestra MG10C para purificar la proteína recombinante GPN empleando como buffer 100 mM Tris-HCl, 100 mM NaCl se descubrió que los contaminantes fueron disminuyendo al ir aumentando la concentración de DTT (figura 28A, líneas 1-4). Al emplear valores menores o iguales a 60 μM (figura 28A, línea 4) todavía se presentaron proteínas no deseadas en la parte superior e inferior de la banda correspondiente a la GPN que fueron disminuyendo en su concentración conforme se aumentó la cantidad empleada de DTT. Para valores de 80 μM y 100 μM (figura 28B, líneas 1, 2) los contaminantes se redujeron de manera significativa. Con la concentración de 100 μM de DTT agregado al buffer (figura 28B, línea 2) se alcanzó una concentración de 2mg/ml de GPN en forma pura medida con el análisis de Bradford.

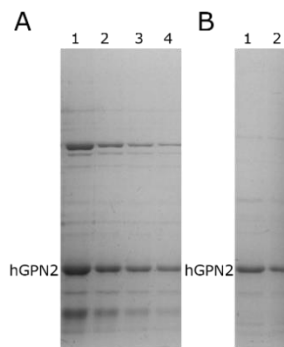


FIGURA 28. ANÁLISIS CON DIFERENTES CONCENTRACIONES DE DTT.

Expresión de la proteína recombinante GPN: A) 10 μM (línea 1), 20 μM (línea 2), 40 μM (línea 3), 60 μM (línea 4). B) 80 μM (línea 1), 100 μM (línea 2). Fuente: Elaboración propia.

4.8. Purificación con aminoácidos

Se empleó el buffer base 100 mM Tris-HCl, 100 mM NaCl a pH 8.2 para preparar nuevas muestras con diferentes concentraciones de Glutamato de

Sodio. Este buffer mostró una disminución de las impurezas conforme se incrementó la concentración del Glutamato de Sodio (figura 29, líneas 1-5). Sin embargo, la cantidad de proteína recombinante GPN también fue disminuyendo como puede apreciarse al reducir el tamaño de la banda correspondiente a la proteína. Después de ser superada la concentración de 100 mM de Glutamato de sodio (figura 29, línea 3) se observó una reducción de la banda de la proteína de interés y las impurezas ya no presentaron disminución manteniendo los mismos contaminantes a las concentraciones de 100, 200 y 300 mM del aminoácido (figura 29, líneas 3-5).

Cuando se sustituyó el Glutamato de Sodio por el aminoácido arginina a 50 mM (figura 29, línea 6) se consiguió el mismo efecto de eliminación de impurezas que con el Glutamato a 100 mM. No obstante, la proteína GPN también sufrió una reducción en el tamaño de la banda lo que implica una disminución en la cantidad de la proteína dentro de la muestra analizada con este buffer y una baja concentración de la proteína de interés.

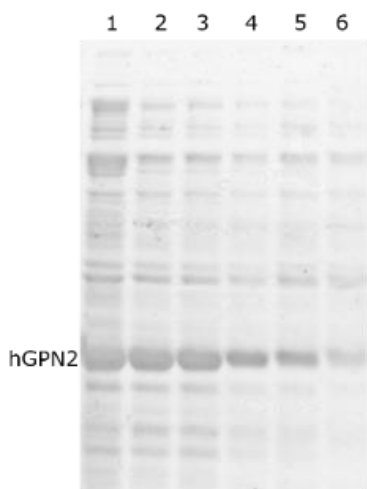


FIGURA 29. ANÁLISIS CON DIFERENTES CONCENTRACIONES DE AMINOÁCIDOS.

Purificación de la proteína recombinante GPN utilizando el buffer 100 mM Tris-HCl, 100 mM NaCl mantenido a pH de 8.2 con Glutamato de Sodio a las siguientes concentraciones: 25 mM (línea 1), 50 mM (línea 2), 100 mM (línea 3), 200 mM (línea 4) y 300 mM (línea 5). Para la línea 6 se sustituyó el Glutamato de Sodio por 50 mM de Arginina. Fuente: Elaboración propia.

Posteriormente, se empleó el mismo buffer y se sustituyó el Glutamato de Sodio por el aminoácido lisina con las concentraciones de 50, 100 y 200 mM (figura 30, líneas 1-3). Al incrementar la concentración del aminoácido se observó una reducción en los contaminantes ubicados en la parte inferior de la proteína GPN, es decir, se redujeron las impurezas que presentan un peso molecular inferior al de la proteína de interés. Incluso también la proteína recombinante expresada sufrió una reducción en el tamaño de su banda. Cuando la concentración utilizada fue superior a los 100 mM de lisina (figura 30, línea 2) los resultados obtenidos en la eliminación de impurezas no fueron mejorados conservando las mismas concentraciones tanto a 200 mM como a 100 mM del aminoácido (figura 30, líneas 2-3).

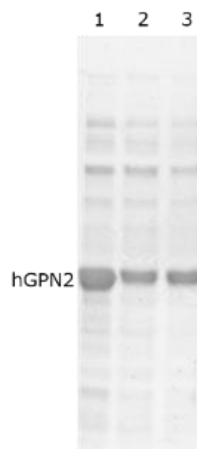


FIGURA 30. ANÁLISIS CON DIFERENTES CONCENTRACIONES DE LISINA.

Purificación de la proteína GPN utilizando el buffer 100 mM Tris-HCl, 100 mM NaCl agregando el aminoácido lisina a las siguientes concentraciones: 50 mM (línea 1), 100 mM (línea 2), 200 mM (línea 3).

Basados en los resultados anteriores se preparó un nuevo buffer con 100 mM Tris-HCl, 100 mM NaCl, 100 mM Glutamato de Sodio, 100 mM Lisina y un pH de 8.2 logrando una reducción bastante notoria de los contaminantes presentes al momento de purificar la proteína empleando la muestra MG10C (figura 31, línea 1). Las proteínas presentes de menor peso molecular casi

fueron eliminadas en su totalidad. Las proteínas de mayor peso molecular, aunque se redujeron, presentaron dos bandas que no pudieron ser reducidas en su totalidad como se aprecia en la figura 31.



FIGURA 31. ANÁLISIS CON BUFFER EMPLEANDO DIFERENTES COMPUESTOS.
Purificación de la proteína GPN empleando el buffer 100 mM Tris-HCl, 100 mM NaCl, 100 mM Lisina, 100 mM Glutamato de Sodio a pH 8.2. Fuente: Elaboración propia.

4.9. Purificación con Buffer Final

Se realizó la purificación de la proteína GPN empleando la muestra MG10C con el buffer final 100 mM Tris-HCl, glicerol al 5%, Tritón X-100 a 0.2%, 100 mM NaCl, 400 μ M EDTA, 100 μ M DTT a un pH de 8.2 que fue construido tomando los mejores resultados de los análisis realizados previamente. El conjunto de aditivos agregados permitió la eliminación total de los contaminantes (las impurezas eliminadas se muestran en la figura 32, línea 1), ya no se presentaron bandas contaminantes en la parte inferior o superior de la proteína GPN por lo que fue concentrada empleando un Amicon® Ultra de 0.5 ml siguiendo las instrucciones del fabricante. Como los contaminantes en cantidades traza también se concentraron (figura 32, línea 2), la muestra se volvió a colocar en la columna IMAC para rotarla durante 20 minutos, se lavó

la columna y se recuperó la proteína después de pasarla por una columna de desalado consiguiendo la proteína GPN completamente pura a una concentración de 30 mg/ml medida mediante el análisis de Bradford (figura 32, línea 3).

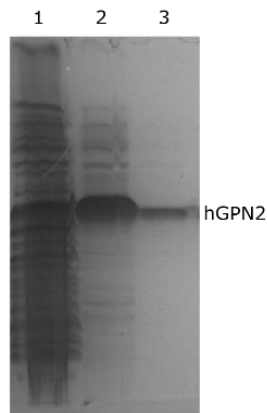


FIGURA 32. CONCENTRACIÓN DE PROTEÍNA PURA.

Purificación de la proteína recombinante GPN. Proteínas contaminantes (línea 1), proteína pura concentrada a 500 µl (línea 2), proteína pura a una concentración de 30 mg/ml (línea 3). Fuente: Elaboración propia.

4.10. Cuantificación de la proteína GPN

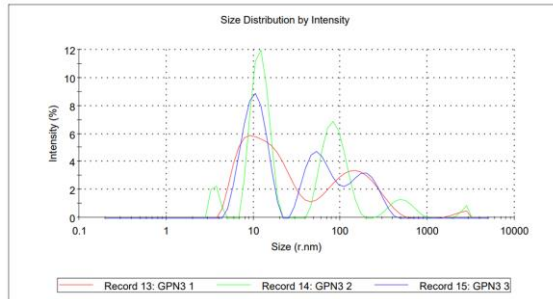
Las cuantificaciones para conocer las concentraciones de las proteínas se realizaron empleando el kit Bradford *protein Assay* de Bio-Rad siguiendo las especificaciones del fabricante. Esta metodología reportó que la concentración máxima obtenida de la proteína GPN purificada empleando el buffer final fue de 30 mg/ml (figura 32, línea 3).

4.11 Análisis de dispersión de luz dinámica

La proteína GPN completamente pura a una concentración de 30 mg/ml, fue analizada en un equipo Malvern Nano S de dispersión de luz dinámica, para determinar el tamaño hidrodinámico de la proteína, conocer la polidispersión de la población proteica y monitorear el fenómeno de agregación de la proteína GPN. El estudio de dispersión dinámica demostró que la proteína de interés tiene un diámetro de 9,366 nm debido a que su radio hidrodinámico tiene un tamaño de 4,683 nm (parámetro *Z-Average*, figura 33B) lo que indica que la proteína no es completamente globular y que algunas regiones quedan fuera del radio hidrodinámico. La polidispersión lograda siguiendo la metodología propuesta proporcionó un valor de 0.124 (parámetro Pdl, figura 33B) lo que indica que la proteína GPN se encuentra monodispersa, homogénea y que puede utilizarse para estudios de cristalografía o de genómica estructural, ya que se encuentra 12.4% polidispersa, lo que es un indicativo de su pureza, es decir, se encuentra 87.6% pura. La bibliografía reporta que con una polidispersión igual o menor al 20% es suficiente para considerarla monodispersa, pura y homogénea. La figura 33A muestra la distribución que tiene la proteína GPN cuando se purifica por el método IMAC empleando buffers tradicionales, indicando que no se encuentra pura, contiene altos contaminantes y una alta polidispersión (Pdl= 0.693, 69.3% polidispersa, indicando una pureza de 30.7%). La figura 33B indica que la muestra es pura, tiene pocos contaminantes y una baja polidispersión.

A

	Size (r.nm):	% Intensity	Width (r.nm):
Z-Average (r.nm): 17.90	Peak 1: 10.57	48.6	3.136
Pdl: 0.693	Peak 2: 61.74	30.5	22.04
Intercept: 0.899	Peak 3: 194.7	20.9	65.73

Result quality: **Good**

B

Results

	Size (r.nm):	% Intensity	Width (r.nm):
Z-Average (r.nm): 4.683	Peak 1: 4.832	95.3	1.330
Pdl: 0.124	Peak 2: 49.38	4.7	14.47
Intercept: 0.957	Peak 3: 0.000	0.0	0.000

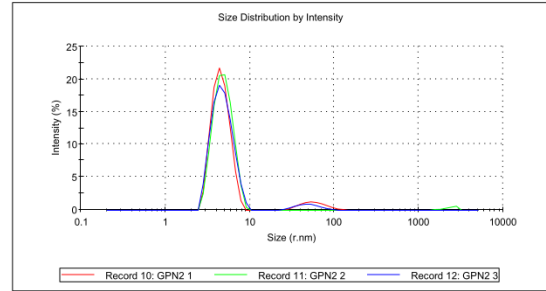
Result quality: **Good**

FIGURA 33. ANÁLISIS DE DISPERSIÓN DE LUZ DINÁMICA DE LA PROTEÍNA GPN.

A) Muestra con alta polidispersión indicando proteína contaminada. B) Muestra con baja polidispersión indicando proteína en forma pura y soluble a una concentración de 30 mg/ml. Fuente: Elaboración propia.

Con los datos de las secciones anteriores se construyeron los geles SDS-PAGE que se emplearon para realizar los análisis de imágenes que se detallan a continuación.

4.12. Análisis preliminar de los geles de proteína GPN a diferentes concentraciones

Los resultados muestran que la proteína recombinante se obtuvo en forma pura por lo que se concentró a diferentes valores (0, 2, 9.5, 14, 14.5, 18, 18.5, 26, 27, 30, mg de proteína por ml de solución) y se agregó al extracto bacteriano tal y como se identifica en las muestras del gel de la figura 34A. Las bandas con mayor grosor corresponden a la GPN con mayor concentración. Se realizaron análisis preliminares de la imagen del gel con muestras de GPN

a diferentes concentraciones para identificar si el estudio general de los perfiles permite detectar la presencia de la proteína o su sobre expresión.

La imagen de la figura 34B muestra el perfil de intensidad del control del peso molecular o *ladder* (carril 1 figura 34A) indicando con los valores mínimos la posición de las bandas que corresponden a las proteínas control empleadas para conocer el peso molecular de las muestras analizadas. El perfil de intensidad permite identificar la posición más no la concentración de las proteínas. Lo anterior ocurre siempre y cuando no existan exceso de proteínas o ruido de fondo como pueden ser proteínas contaminantes.

La figura 34C contiene el perfil de intensidad del carril 4 (figura 34A) donde puede apreciarse que el pico de menor tamaño (mínimo) indica la presencia de la proteína con mayor concentración en el carril. Sin embargo, no permite encontrar su concentración ni la relación con el resto de las proteínas presentes en la misma muestra (mismo carril) ni su peso molecular. En este análisis tampoco pueden detectarse la presencia del resto de proteínas expresadas ya que solo aparecen como si fuera ruido de fondo y solo se identifica una sola proteína.

Finalmente, figura 34D muestra el perfil de intensidad de las regiones que contienen la proteína GPN a diferentes concentraciones que corresponden a los carriles 2 al 11 de la figura 34A. Puede apreciarse el ruido de fondo generado por las diferentes cantidades de proteínas que incluye el extracto total y por las diferentes concentraciones de GPN que presentan las muestras, no puede distinguirse el grosor o altura de los picos para relacionarlos con cada una de las diferentes sobre expresiones de la proteína recombinante. Estos resultados indican que los métodos tradicionales no permiten identificar la presencia de una banda de proteína específica y por lo tanto es necesario aplicar diferentes métodos que permitan extraer sus características.

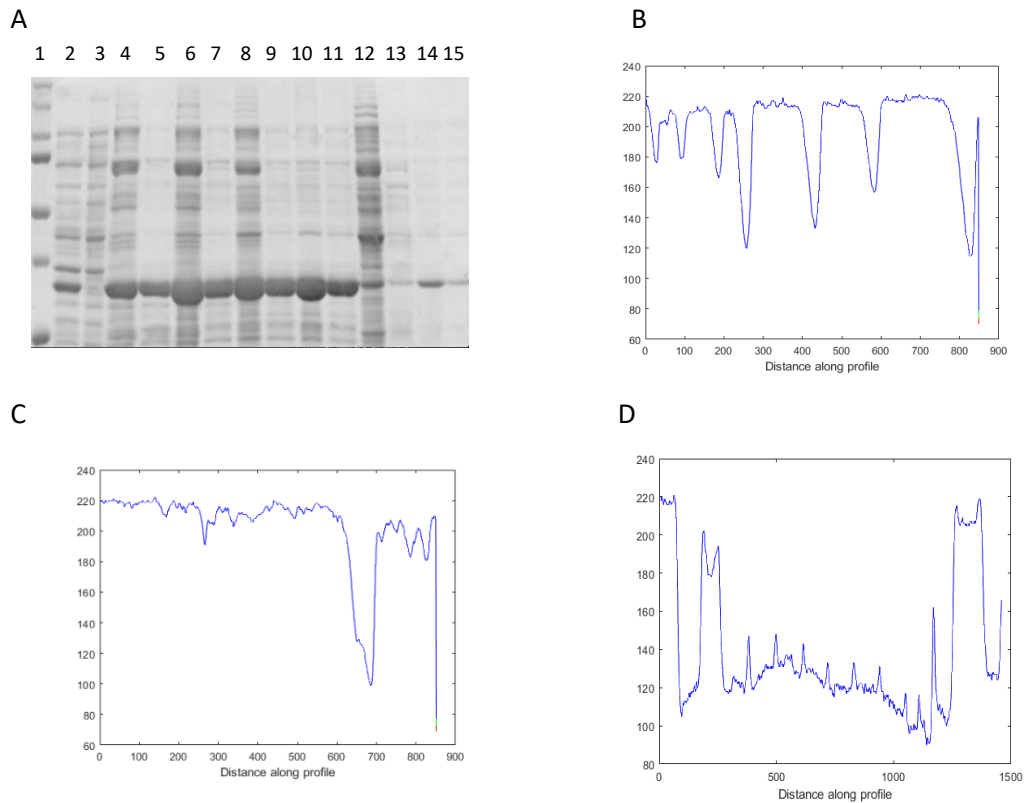


FIGURA 34. ANÁLISIS PRELIMINAR DE LAS IMÁGENES DE GELES SDS-PAGE.

A) Gel SDS-PAGE que contiene proteína GPN expresada a diferentes concentraciones. Carril 1, control de peso molecular; carriles 2-11 GPN concentraciones de 2, 0, 18.5, 9.5, 30, 18, 27, 14.5, 26, 14 $\mu\text{g/ml}$, respectivamente. B) Perfil de intensidad del carril 1 del gel de la figura 34A. C) perfil de intensidad del carril 4, figura 34A. D) perfil de intensidad de los carriles 2-11 de la figura 34A que incluye solamente a la proteína recombinante GPN. Fuente: Elaboración propia.

4.13 Preprocesamiento y Extracción de características empleando metodología desarrollada (perfil de imagen basado en segmentación de imágenes binarias empleando una máscara binaria)

Las imágenes de los geles de poliacrilamida fueron ajustadas a un tamaño de 600 x 400 píxeles previo al análisis con la metodología propuesta.

Se analizó cada gel empleando una máscara binaria del tamaño de un píxel horizontal x 400 pixeles verticales. La función de la máscara fue aislar el resto del gel y permitir solamente el paso de los pixeles incluidos en una matriz de 400 x 1. Se calculó su histograma, se aplicó un método de binarización (como Niblack, Sauvola, *thresholding* u Otsu) y se segmentó para separar de la imagen binaria los valores de intensidad de los pixeles blancos y negros. Se eliminaron todos los datos que pertenecen al histograma y se eligió solamente el valor que le corresponde a la máxima intensidad del blanco, con el fin de que al ir disminuyendo y se acerque a cero se pueda utilizar para graficar un nuevo perfil de la imagen que identifique la separación entre las líneas o las bandas del gel. A esta nueva metodología se le denominó *Perfil de Imagen basado en segmentación de imágenes binarias* y se detalla a continuación:

La imagen del gel obtenida después de la tinción con azul de Coomasie fue convertida a escala de grises para ser binarizada, se erosionó para incrementar el espacio existente entre muestras y se le aplicó la máscara binaria (ver figura 35).

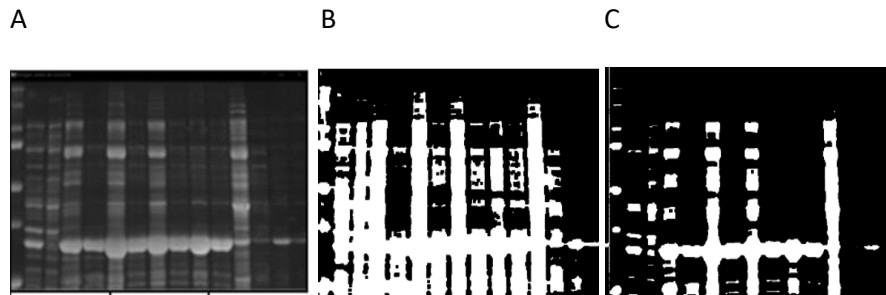


FIGURA 35. PREPROCESAMIENTO DE LAS IMÁGENES DE GELES.

A) Imagen en escala de grises. B) Imagen binarizada y C) imagen erosionada. Fuente: Elaboración propia.

A la imagen resultante (figura 35C) se le aplicó una máscara binaria que consiste en una matriz de 1x400 pixeles (por cuestiones de visualización, las imágenes de la figura 36 presentan una máscara de 10x800 pixeles). La

imagen 36C muestra la región del gel que fue analizada por la máscara y a la que se le aplicó la binarización y segmentación mediante el método de Otsu.

Cuando se calculó el histograma en la región del gel que se encuentra cubierta por la máscara binaria, se obtuvo el patrón de la gráfica 36D que incluyó la distribución de los píxeles. En cambio, al aplicar la segmentación en la imagen binaria por el método de Otsu en la región contenida dentro de la máscara se consiguió el histograma que se muestra en la gráfica de la figura 36E, que solamente grafica el valor de máxima intensidad del color negro (píxel 0) y del color blanco (píxel 255).

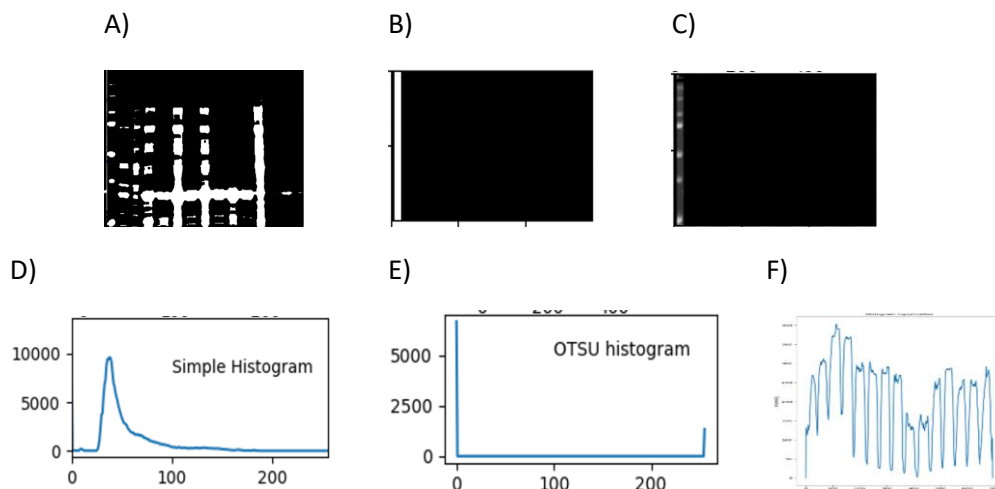


FIGURA 36. PERFIL DE IMAGEN BASADO EN SEGMENTACIÓN DE IMÁGENES BINARIAS.

A) Imagen previa a la que se aplicó la máscara binaria. B) Representación de la máscara binaria que corresponde a un tamaño de 1x400 píxeles. C) Segmento del gel SDS-PAGE que fue analizado por la máscara binaria. D) Histograma simple de la imagen C. E) Histograma de la imagen C después de ser binarizada, erosionada y segmentada por el método de Otsu. F) Perfil obtenido al graficar los valores del arreglo que contienen solamente los píxeles blancos obtenidos por la máscara binaria. Fuente: Elaboración propia.

Conforme la máscara fue avanzando por los 600 píxeles que contiene la imagen, se fueron almacenando dentro de un arreglo solamente los valores obtenidos por los píxeles blancos (posición 255 del histograma). Estos valores

almacenados fueron graficados creando un nuevo perfil (perfil de imagen basado en segmentación de imágenes binarias, ver figura 36F) y se usó para realizar el análisis de las imágenes de los geles de proteínas SDS-PAGE.

4.14. Detección de proteína sobre expresada en geles SDS-PAGE empleando Perfil de Imagen basado en segmentación de imágenes binarias

Se aplicó el perfil de imagen basado en segmentación de imágenes binarias para la búsqueda de la sobre expresión de la proteína GPN en el gel mostrado en la figura 37A. La imagen se analizó de forma ecualizada (figura 37B) y sin ecualizar (figura 37C).

La metodología requirió de almacenar en un arreglo los valores de intensidad máximos blancos obtenidos al pasar la máscara binaria a través de la imagen para ser graficados posteriormente (figuras 37B y 37C). Las gráficas presentaron el mismo patrón con la diferencia que la ecualizada mantuvo picos más altos (figura 37B) que la no ecualizada (figura 37C). La diferencia se presentó en los mínimos, la imagen no ecualizada grafica los mínimos más cerca de la línea base o cero por lo que puede utilizarse para buscar un umbral que contenga todos los picos y de forma semi-automática encontrar la detección del número de muestras presentes en el gel. Por ello, para los análisis posteriores se utilizó la imagen sin ecualizar en el perfil de imagen basado en segmentación de imágenes binarias.

Los máximos de la gráfica indicaron las zonas donde se tiene una región con mayor color blanco lo que proporcionó una relación directa con la cantidad de GPN expresada en las muestras. Los carriles 5 al 15 de la figura 37A mostraron un extracto con la misma cantidad de proteínas endógenas a las que se les han agregado diferentes concentraciones de proteína

recombinante, el tamaño de las manchas indicó una mayor concentración de GPN en los carriles 5, 6, 11 y 12, y una menor concentración en las líneas 9 y 10. Estos valores fueron detectados en la gráfica, donde se tuvieron los picos de mayor tamaño en 5, 6, 11 y 12 y los picos de menor altura representaron las bandas con menor concentración de la proteína recombinante, es decir, la menor sobre expresión (figura 37B y 37C).

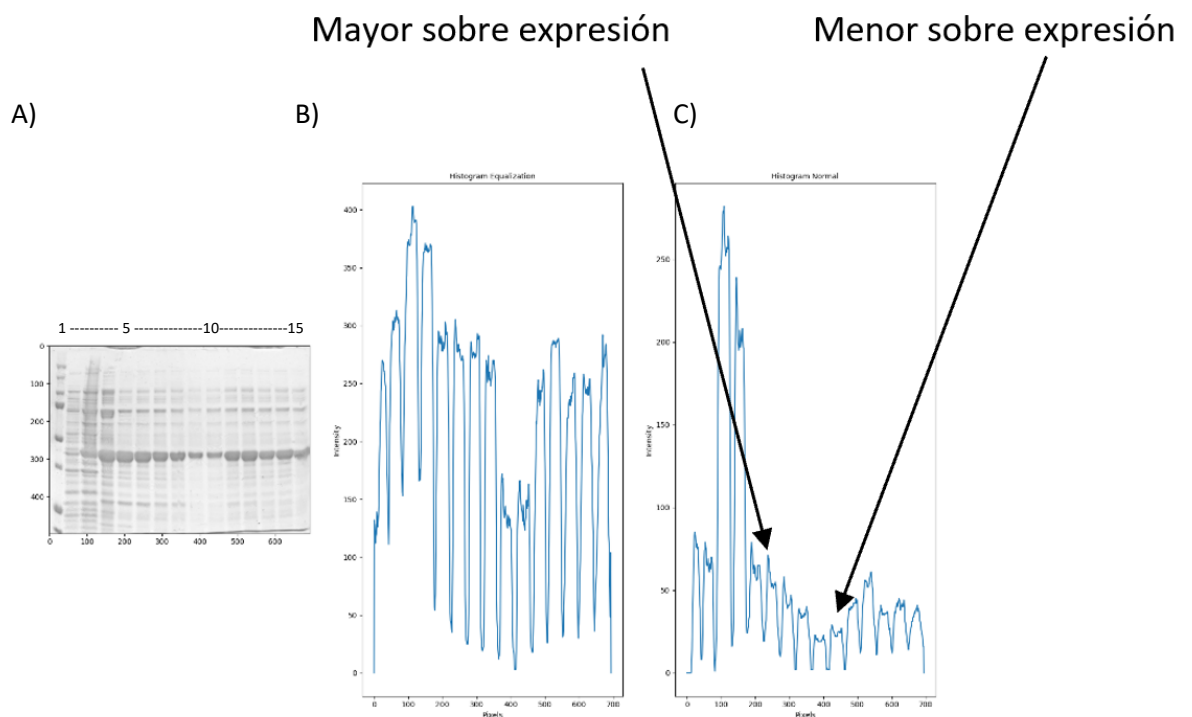


FIGURA 37. ANÁLISIS DEL NUEVO PERFIL.

A) gel que incluye en el carril 1 el buffer de carga con los pesos moleculares control; carril 2, 3 y 4 extracto total de proteínas; carriles 5 al 15 proteína GPN con diferentes niveles de concentración. B) gráfica obtenida aplicando el perfil de imagen basado en segmentación de imágenes binarias sobre la figura 37A habiendo previamente ecualizado la imagen. C) gráfica obtenida aplicando la misma metodología sobre la imagen 37A sin ecualizarla. Fuente: Elaboración propia.

Los valores de cada uno de los máximos de la gráfica de la figura 37C se muestran en la tabla 3. En ella se pudo identificar que el valor máximo se tiene en el carril 5 representando la mayor sobre expresión de la proteína

recombinante en esta muestra y las menores expresiones se consiguen en los carriles 9 y 10 con unos valores máximos registrados como 23 y 29.

TABLA 3. DATOS DEL NUEVO PERFIL.

Valores de los máximos de los picos obtenidos de la gráfica de la figura 37C. Fuente: Elaboración propia.

Lane	5	6	7	8	9	10	11	12	13	14	15
Valor	79	71	58	40	23	29	45	61	41	45	41

Para identificar la funcionalidad del método se buscó expresar una concentración controlada de proteína, para ello se concentró la proteína BSA de 0.5 mg/ml a 1 mg/ml y 2 mg/ml y se realizó la electroforesis. El gel obtenido se muestra en la figura 38A donde se observó el incremento en el ancho de la banda (o mancha) al momento de concentrar la proteína. Se realizó el análisis de la imagen con el método de perfil de imagen basado en segmentación de imágenes binarias y se obtuvo la gráfica de la figura 38B, donde puede apreciarse que el pico máximo corresponde a la proteína BSA de mayor concentración (2 mg/ml) y que el menor de los máximos locales pertenece a la menor concentración de BSA (0.5 mg/ml).

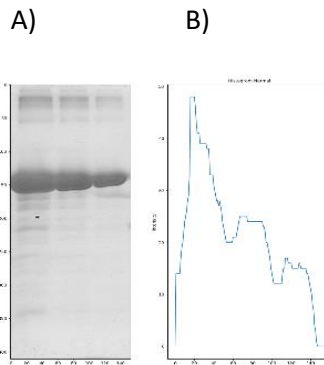



FIGURA 38. ANÁLISIS EMPLEANDO LA PROTEÍNA BSA.

A) Gel SDS-PAGE de la proteína BSA concentrada. B) Gráfica de los valores promediados individuales a una concentración de 2 mg/ml (carril 1), 1 mg/ml (carril 2), y 0.5 mg/ml (carril 3). Nótese los máximos locales de las proteínas sobre expresadas. Fuente: Elaboración propia.

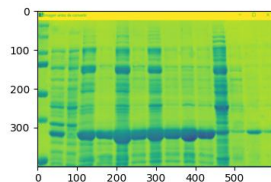
Para probar la eficacia del método, se prepararon las concentraciones de la proteína GPN con valores conocidos (tabla 4) agregando diferentes cantidades de extractos celulares para incluir distintas proteínas e incrementar el ruido o background de fondo.

TABLA 4. PROTEÍNA PURA A DIFERENTES CONCENTRACIONES.

Muestras con diferentes concentraciones de proteína recombinante GPN distribuidas en el gel de poliacrilamida. Fuente: Elaboración propia.

Línea	2	3	4	5	6	7	8	9	10	11
Concentración mg/ml	2	0	18.5	9.5	30	18	27	14.5	26	14
Gel										

A)



B)

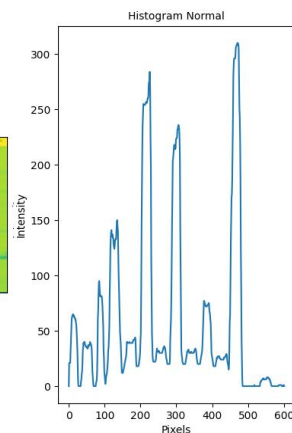


FIGURA 39. PERFIL OBTENIDO DEL GEL COMPLETO.

A) Gel SDS-PAGE con muestras a diferentes concentraciones de GPN. B) gráfica obtenida empleando el perfil de imagen basado en segmentación de imágenes binarias de 39A. Fuente: Elaboración propia.

Se aplicó la metodología propuesta en el gel con las muestras de GPN a diferentes concentraciones (ver figura 39A y B). Los resultados obtenidos

indican que las muestras con mayor concentración de proteínas tienen los máximos locales con valores más grandes. Sin embargo, al comparar el carril 4 con el carril 10 la gráfica mostró que se tiene una mayor concentración de proteínas en el 4 (con un máximo de 150), lo que no coincide con las concentraciones preparadas, ya que la línea 4 tiene una concentración de 18.5 mg/ml y la línea 10 un valor de 26 mg/ml (máximo de 75). Esta inconsistencia en los resultados se debe a que el carril 4 presentó una mayor cantidad de proteínas en el extracto total comparado con las proteínas de la muestra 10. Esto permitió corroborar que el perfil de imagen basado en segmentación de imágenes binarias solamente es consistente cuando el ruido de fondo disminuye o es el mismo en todas las muestras, es decir, se debe tener la misma cantidad de proteínas en el extracto total para poder detectar si la proteína de interés se encuentra sobre expresada o no. Estas proteínas que no son de interés se encontraron en la parte superior e inferior de la proteína recombinante (GPN) y pueden considerarse como contaminantes, por lo tanto, deben ser eliminadas.

Para conseguir su eliminación se buscó el peso molecular de las proteínas para identificar la proteína de interés, se separó del resto y se repitió el análisis tomando en cuenta la proteína recombinante y finalmente se volvió a detectar su sobre expresión.

Tomando en cuenta que el control de peso molecular no tiene ruido de fondo, se seleccionó y se separó del resto del gel (ver figura 40A y línea 1 de figura 40B), se le aplicó el perfil de imagen basado en segmentación de imágenes binarias y se obtuvieron las gráficas que se muestran en la figura 41B y 41C.

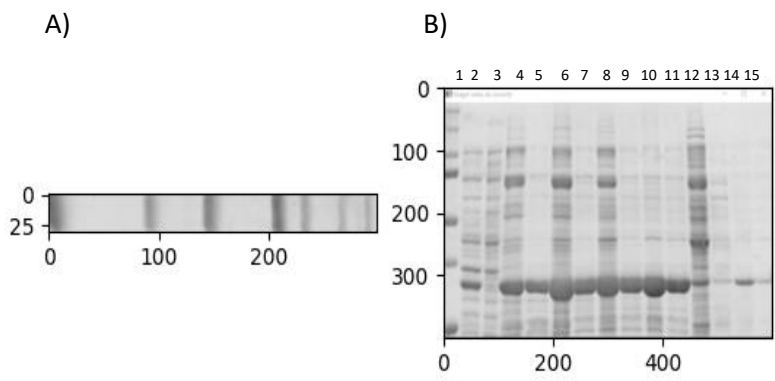


FIGURA 40. SEPARACIÓN DE LA MUESTRA CONTROL.

A) control de peso molecular obtenido del primer carril de B. B) Gel completo de donde se obtiene la muestra A. Fuente: Elaboración propia.

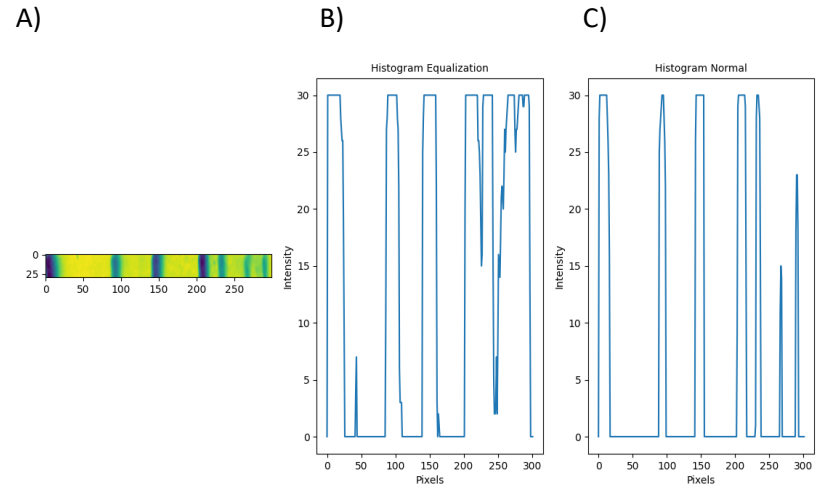


FIGURA 41. PERFIL OBTENIDO SOBRE LA MUESTRA CONTROL.

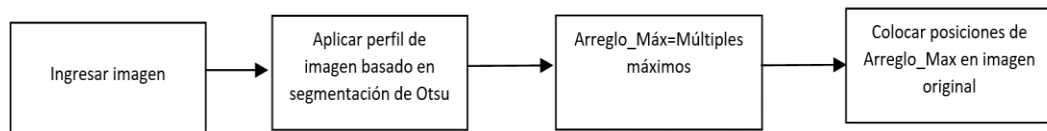
A) muestra que corresponde al carril 1 del gel de la figura 40B y contiene los pesos moleculares conocidos. B) gráfica obtenida después de ecualizar la imagen 41A y realizar el perfil de imagen basado en segmentación de imágenes binarias. C) gráfica de la imagen 41A sin ecualizar aplicando el PIBSIB. Fuente: Elaboración propia.

La imagen procesada (figura 41A) fue analizada de forma ecualizada (figura 41B) y sin ecualizar (figura 41C), estos resultados mostraron que cuando no existe ruido de fondo es mejor no realizar la ecualización de la imagen ya que esto incrementó ruido en la gráfica al momento de aplicar la metodología (figura 41B). En cambio, la gráfica obtenida sin realizar la ecualización de la imagen permitió detectar perfectamente los máximos que corresponden a los

valores donde se encuentran las bandas de control de peso molecular (figura 41C).

Se utilizó un arreglo para almacenar los valores máximos múltiples obtenidos al emplear el perfil de imagen basado en segmentación de imágenes binarias (PIBSIB) y se relacionaron con la posición de las proteínas del carril 1 de la figura 40B (ver figura 42A y 42B). Estos valores fueron asignados a cada una de las posiciones de las proteínas dentro de la imagen y se dibujaron para verificar que correspondieran los pesos moleculares reales (de la imagen) con los pesos moleculares calculados por los máximos conseguidos por el PIBSIB (ver figura 43).

A)



B)

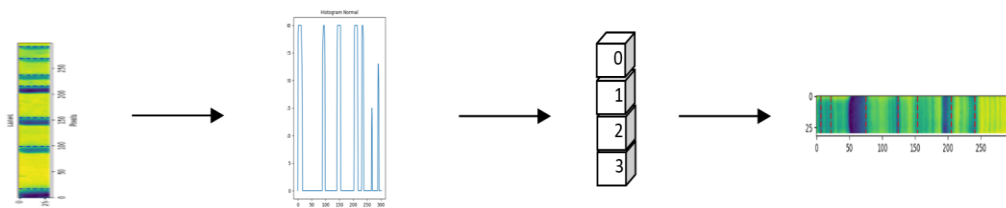


FIGURA 42. DETECCIÓN DE BANDAS.

A) Diagrama que relaciona el peso molecular de la muestra control con los máximos obtenidos por el perfil de imagen basado en segmentación de imágenes binarias para encontrar la posición de las proteínas. B) Esquema del proceso descrito en 42A. Fuente: Elaboración propia.

Los pesos moleculares fueron proporcionados por el fabricante y se ingresaron previamente al programa. Se utilizó la relación con los máximos obtenidos para aplicar diferentes métodos de interpolación (lineal, cúbica y *nearest*) y se eligió el de menor margen de error para tener una ecuación que permitió predecir

los pesos futuros de las proteínas. La ecuación determinada se empleó para detectar el peso molecular y la posición de la proteína GPN dentro del gel SDS-PAGE (ver figura 44). Los valores obtenidos por cada uno de los métodos de interpolación utilizados se muestran en la tabla 5.

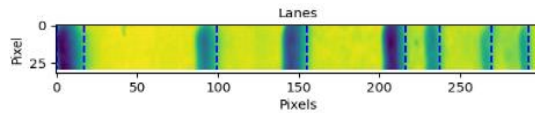


FIGURA 43. CORRELACIÓN ENTRE BANDAS DETECTADAS AUTOMÁTICAMENTE Y LA IMAGEN DE LA MUESTRA CONTROL.

Bandas detectadas por la metodología propuesta. Las bandas están marcadas en azul dentro de la muestra del gel SDS-PAGE identificando la posición de las proteínas control. Fuente: Elaboración propia.

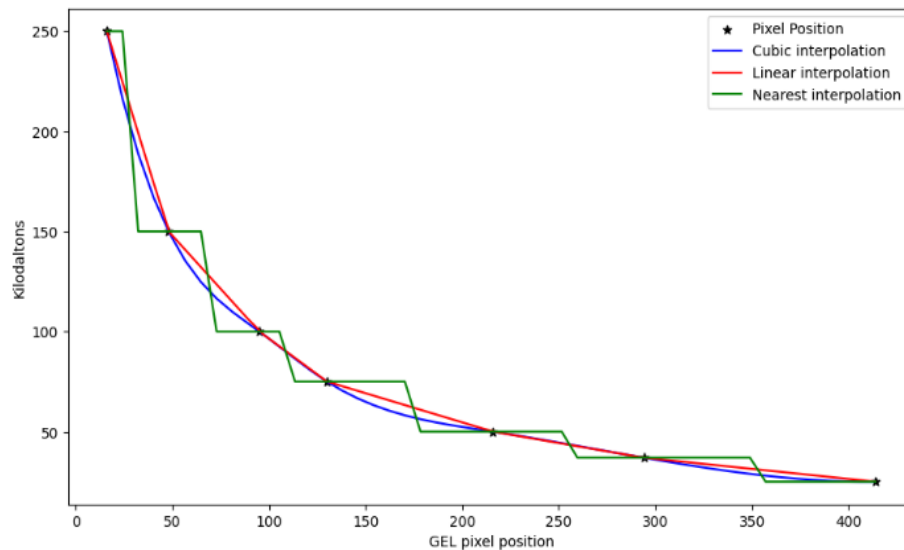


FIGURA 44. GRÁFICA PARA DETECTAR PESO MOLECULAR.

Métodos de interpolación empleados para predecir el peso molecular de las proteínas analizadas. Eje X corresponde a la posición de los pixeles de las bandas de proteína dentro de la imagen del gel SDS-PAGE y el eje Y son los pesos moleculares asignados. Fuente: Elaboración propia.

TABLA 5. MARGEN DE ERROR EN MÉTODOS DE INTERPOLACIÓN.

Error obtenido al aplicar tres métodos numéricos de interpolación al Ladder (carril 1) del gel de poliacrilamida. Fuente: Elaboración propia.

Métodos de interpolación	Peso calculado (KDa)	Error total %
Lineal	33.4	3.35648148
Nearest	37.0	7.060185185
Cúbica	31.38	9.194960019

Los resultados obtenidos en la tabla 5 indicaron que el método de interpolación lineal es el que menos error reportó con un valor de 3.35% y se acercó más al peso molecular real de la proteína GPN con un valor de 34.56 Kilo Daltones (KDa), por lo tanto, la fórmula que se utilizó para identificar el peso molecular del resto de las proteínas en cualquiera de las muestras analizadas quedó definida como:

$$y(x) = y_i + \frac{(y_{i+1}-y_i)(x-x_i)}{(x_{i+1}-x_i)} \quad (1)$$

Donde $0 \leq y \leq 400$ corresponde al intervalo que contiene el número de pixeles de la imagen y $0 \leq x \leq 250$ corresponde al peso molecular de las proteínas.

Se utilizó el perfil de imagen basado en segmentación de imágenes binarias para realizar la detección automática de todas las muestras que contiene el gel. Ya que los máximos permiten interpretar la mayor intensidad del color blanco, entonces los mínimos deberán detectar la ausencia de este, por lo que pueden relacionarse con las zonas donde las muestras se separan o donde no hay proteínas. Por ello, se realizó una correspondencia entre los mínimos obtenidos con el PIBSIB y la separación de las muestras que contiene el gel de poliacrilamida

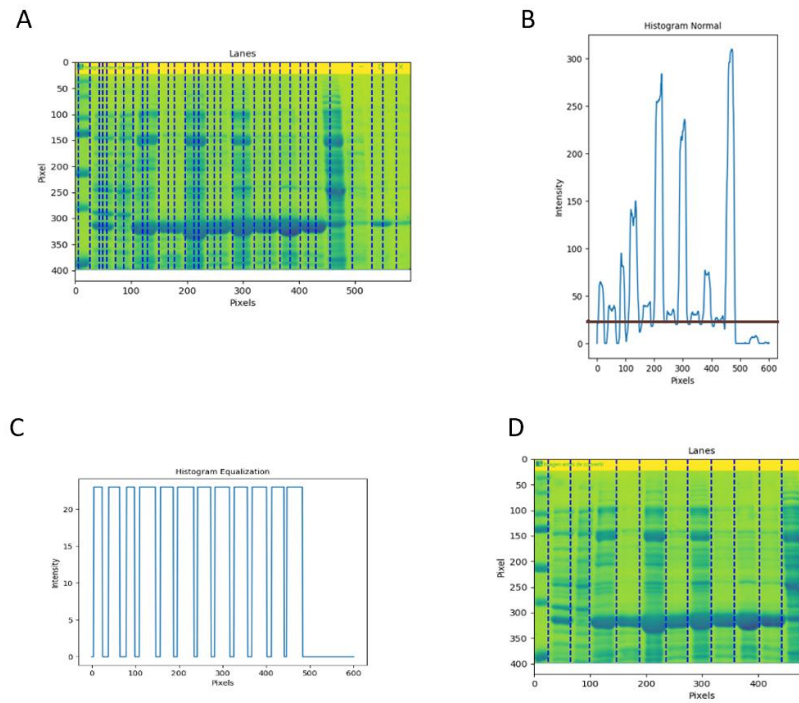


FIGURA 45. CÁLCULO DE UMBRAL.

A) Múltiples máximos detectados. B) Región de corte elegida como umbral que incluyó todos los mínimos. C) Gráfica obtenida después de aplicar el umbral de la figura 45B. D) Total de muestras detectadas de forma automática después de aplicar el umbral de la figura 45B. Fuente: Elaboración propia.

Los máximos múltiples detectados son generados por el ruido de fondo o background que existe entre las proteínas (figura 45A) y son eliminados al elegir un umbral que contenga todos los mínimos y elimine los múltiples máximos a manera de filtro pasa bajas (línea negra en la figura 45B). Después de aplicar el filtro se obtuvo una nueva gráfica que ya no incluye los máximos múltiples y cuyos mínimos corresponden a la separación entre las líneas del gel (figura 45C). El método desarrollado permite elegir las regiones que contienen la menor cantidad de pixeles blancos y las relaciona con las zonas donde se encuentran las muestras consiguiendo detectar los experimentos presentes en el gel. Los valores encontrados fueron indicados mediante líneas azules dentro de la imagen original para verificar la detección automática de todas las muestras (figura 45D) y los datos resultantes fueron almacenados en

un arreglo para ser utilizados posteriormente ya que contienen las posiciones del gel donde se separan todas y cada una de las muestras.

Se eligió una muestra aleatoria en el gel (figura 46) para repetir el procedimiento de detección de peso molecular poniendo énfasis en la proteína recombinante GPN con un peso molecular de aproximadamente 34.56 KDa. El perfil de imagen basado en segmentación de imágenes binarias pudo detectar de forma automática la banda de mayor grosor, primero se identificaron las posiciones de las proteínas de mayor concentración (ver figura 47A) y posteriormente se identificó de forma automática la proteína GPN tomando en cuenta su peso molecular como lo muestra la figura 47B.

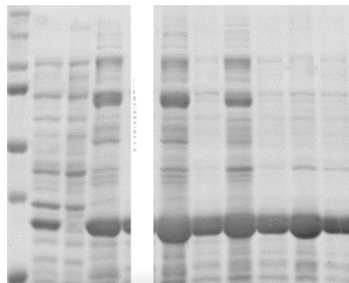


FIGURA 46. MUESTRA ALEATORIA PARA ANÁLISIS.

Elección de una muestra elegida de forma aleatoria dentro del gel SDS-PAGE que contiene diferentes concentraciones de la proteína GPN. Fuente: Elaboración propia.

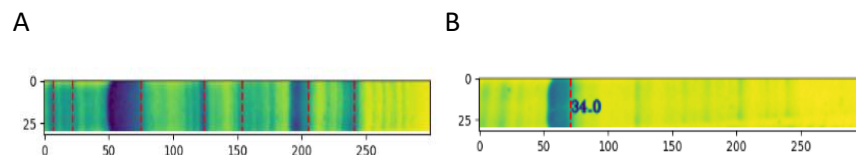


FIGURA 47. DETECCIÓN DE GPN EN BASE A SU PESO MOLECULAR.

A) Bandas detectadas de forma automática por el perfil de imagen basado en segmentación de imágenes binarias. B) Peso molecular detectado para la proteína GPN. Fuente: Elaboración propia.

Habiendo identificado previamente en un arreglo las posiciones que separan las muestras incluidas en el gel y conociendo el peso molecular de la proteína GPN (34 KDa) se eligió como región de interés (ROI) la zona que comprende la expresión de la proteína recombinante a diferentes concentraciones como

lo muestra la figura 48 y se utilizó para realizar una comparación de la sobre expresión de la proteína e identificar si la metodología propuesta permite detectar aquellas proteínas que se encuentran con mayor o menor sobre expresión.

Los resultados del perfil de imagen basado en segmentación de imágenes binarias obtenidos en la zona catalogada como ROI se muestran en la figura 49. Los valores máximos obtenidos de cada una de las muestras se incluyen en la tabla 6 (perfil de imagen basado en segmentación de imágenes binarias aplicado a la zona ROI que contiene solamente la proteína GPN a diferentes concentraciones). Los datos identificaron que la menor sobre expresión la tuvo la muestra de la línea 2 (con un valor máximo de 8.8 y una concentración de 2 mg/ml) seguida por orden de sobre expresión de la línea 5 (17.08889, 9.5 mg/ml), línea 11 (18.13333, 14 mg/ml), línea 4 (18.466667, 14.5 mg/ml), línea 9 (21.288889, 18 mg/ml), línea 7 (21.33333, 18.5 mg/ml), línea 10 (25.044445, 26 mg/ml), línea 8 (26.177778, 27 mg/ml) y línea 6 (26.688889, 30 mg/ml). Estos resultados representaron correctamente la sobre expresión, lo que no ocurrió cuando se analizó previamente el gel completo empleando el PIBSIB (tabla 6, metodología propuesta, gel completo).

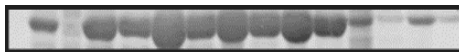


FIGURA 48. REGIÓN DE INTERÉS (ROI).

Elección de la zona de interés o ROI que incluye las bandas de proteína GPN a diferentes concentraciones. Fuente: Elaboración propia.

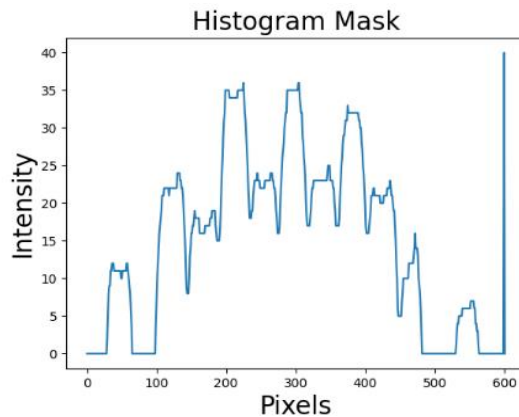


FIGURA 49. PERFIL DE ZONA ROI.

Gráfica de intensidad de blancos empleando la máscara binaria donde los máximos indican la expresión de la proteína del gel de la figura 48. Fuente: Elaboración propia.

Estos resultados indicaron que el perfil de imagen basado en segmentación de imágenes binarias desarrollado funcionó perfectamente para detectar la sobre expresión de las proteínas siempre y cuando el ruido de fondo sea mínimo, casi nulo o tenga las mismas concentraciones en todas las muestras como ocurre en el caso de que todas las muestras fueran del mismo paciente en diferentes periodos temporales. En estas condiciones, la metodología propuesta permitió detectar cual proteína se encontró más sobre expresada e incluso proporcionó el orden de sobre expresión.

Se aplicó el perfil de imagen basado en segmentación de imágenes binarias empleando 44 geles conteniendo 669 muestras donde los geles SDS-PAGE conservaron las mismas características que incluyen el mismo color y ningún defecto como puede ser gel roto o distorsionado por haber sido mal preparado.

Para medir la exactitud se empleó la matriz de confusión definiendo los verdaderos positivos (TP) cuando la línea o banda de proteína existe y la metodología propuesta lo encuentra, los falsos negativos (FN) cuando la línea existe y no se encuentra, los falsos positivos (FP) si la línea no existe y se encuentra, y los verdaderos negativos (TN) cuando la línea no existe y no se encuentra. La matriz de confusión obtenida después de analizar las 669

muestras se indica en la tabla 7. La exactitud obtenida fue de 0.985052 (ver tabla 9, exactitud geles homogéneos).

Se repitió el análisis empleando el perfil de imagen basado en segmentación de imágenes binarias usando 90 geles lo que hizo un estudio de 1561 muestras analizadas para buscar la sobre expresión.

Para medir la exactitud se utilizó la matriz de confusión empleando los mismos parámetros que los geles homogéneos (TP, FN, FP, TN). La matriz de confusión obtenida después de analizar las 1561 muestras se indica en la tabla 8. La exactitud obtenida fue de 0.91736 (ver tabla 9, exactitud geles heterogéneos).

TABLA 6. COMPARACIÓN DE RESULTADOS APLICANDO DIFERENTES METODOLOGÍAS.

Muestras con diferentes concentraciones de proteína recombinante GPN distribuidas en el gel de poliacrilamida. Fuente: Elaboración propia.


Línea	2	3	4	5	6	7	8	9	10	11
[1]	2	0	18.5	9.5	30	18	27	14.5	26	14
[2]	513.6	0	830.2	680.1	1373.8	830	1061.5	934.5	983.5	869.25
[3]	206	0	293	60	333	141	301	133	183	173
[4]	8.8	0	18.5	17.1	26.7	21.3	26.2	21.3	25	18.1
[5]	495.5	0	969.6	933	1457.2	1132.2	1336.2	1134.9	1422.1	993.1
[6]	535.8	0	1189.8	1023.4	1646.8	1318.2	1557.1	1269.1	1585.9	1174.6
[7]	11.3	0	27.3	22.5	38.5	30.7	36.8	30.3	36.1	27.1
[8]	19	0	35	27	46	35.8	43	35.5	44	32
Gel										
<p>[1] Concentración mg/ml [2] Manual (squares) [3] Metodología propuesta (gel completo) [4] Metodología propuesta (ROI con GPN) [5] Área con segmentación K-Means [6] Área con segmentación Otsu [7] Segmentación semántica (ROI con GPN) [8] Segmentación semántica (Gel completo)</p>										

TABLA 7. MATRIZ DE CONFUSIÓN DE MUESTRAS HOMOGÉNEAS.

Matriz de confusión obtenida al analizar 669 muestras con la proteína GPN expresada a diferentes concentraciones. Fuente: Elaboración propia.

		Predichas	
		Positivas	Negativas
Reales	Positivas	TP=310	FN=8
	Negativas	FP=2	TN=349

TABLA 8. MATRIZ DE CONFUSIÓN DE MUESTRAS HETEROGÉNEAS.

Matriz de confusión obtenida al analizar 1561 muestras con la proteína GPN expresada a diferentes concentraciones. Fuente: Elaboración propia.

		Predichas	
		Positivas	Negativas
Reales	Positivas	TP=671	FN=105
	Negativas	FP=24	TN=761

TABLA 9. EXACTITUD OBTENIDA DE GELES HOMOGÉNEOS Y HETEROGÉNEOS.

Resultados de exactitud obtenidos de la matriz de confusión empleando el programa Matlab R2022a. Fuente: Elaboración propia.

Datos de exactitud obtenida		
Exactitud de geles homogéneos	Exactitud geles heterogéneos	Exactitud por segmentación semántica geles heterogéneos
0.985052	0.91736	0.95085 (etiquetas: banda, fondo)

Se comparó si el área de las bandas puede ser detectada por otros métodos para calcular el nivel de sobre expresión de la proteína GPN, por lo que se emplearon tres técnicas diferentes para realizarlo. Ya que no existe ninguna metodología que permita analizar los geles completos o muestras que incluyan diferentes proteínas, se recortaron las bandas correspondientes a la proteína GPN a diferentes concentraciones y se midieron las áreas de forma independiente empleando rectángulos, se utilizó segmentación por *K-means* y finalmente segmentación por el método de Otsu (segmentación de Otsu) utilizando el programa Matlab R2022a (ver figura 50).

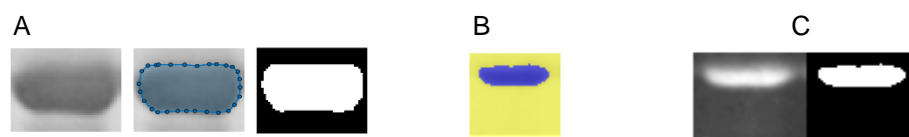


FIGURA 50. ÁREAS OBTENIDAS DE MUESTRA GPN EMPLEANDO DIFERENTES METODOLOGÍAS.
 A) Cálculo manual de áreas de las bandas. B) Áreas calculadas por segmentación de *K-means*. C) Áreas calculadas por segmentación de Otsu. Fuente: Elaboración propia.

Para identificar si el cálculo de área permite encontrar la sobre expresión de las proteínas, los datos obtenidos de las áreas con todos los métodos utilizados se normalizaron y se graficaron (ver figura 51), los datos se colocaron en la tabla 10.

Para su análisis, se realizó la siguiente medición:

Sea x_n el valor *n-ésimo* del área medida y x_{n+1} el valor medido de la banda con mayor concentración de x_n .

Si $x_{n+1} > x_n \Rightarrow x_{n+1} - x_n > 0$, por otro lado, si $x_{n+1} < x_n \Rightarrow x_{n+1} - x_n < 0$

Como se observó en la tabla 10 y en la figura 51, el resultado del análisis anterior aplicado a cada una de las medidas indicó que el método manual presentó tres valores negativos (en la línea 5, línea 7 y línea 9 de la tabla 10 y figura 51), el GPN-ROI ninguno, la segmentación *K-means* tres valores negativos (línea 5, línea 7 y línea 9 de la tabla 10 y figura 51) y el Método de Otsu un valor negativo (línea 9, tabla 10 y figura 51).

Los resultados muestran que el perfil de imagen basado en segmentación de imágenes binarias desarrollado permitió encontrar la sobre expresión de forma correcta superando a otros métodos como el manual, segmentación de *K-means* y segmentación de Otsu. Los que más fallaron al momento de medir el área de la proteína sobre expresada fueron el método manual y el de *K-means*.

El perfil de imagen basado en segmentación de imágenes binarias consiguió encontrar la cantidad total de muestras presentes en el gel y utilizó el arreglo generado para conocer la separación o región de corte para el análisis posterior de la proteína de interés. Identificó el peso molecular de la proteína recombinante para generar la región de interés conocida como GPN-ROI y calculó las áreas de la GPN de forma correcta logrando identificar la sobre expresión ordenándolas de menor a mayor sin tener que analizarlas por separado. Por otro lado, el resto de las técnicas analizadas (manual, segmentación *K-means* y segmentación de Otsu) presentaron errores al momento de identificar la sobre expresión de la proteína recombinante. Para poder realizar el cálculo fue indispensable separar cada una de las muestras de forma manual para obtener pequeñas regiones que contuvieran la proteína y pudiera realizarse la medición de la segmentación ya que hasta el momento no existe técnica que permita analizar el gel completo.

TABLA 10. NORMALIZACIÓN DE RESULTADOS OBTENIDOS.

Normalización de la información empleando los métodos: manual, GPN-ROI, segmentación de *K-means* y segmentación de Otsu de la tabla 6. Fuente: Elaboración propia.

Línea	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
1	0	0	0	0	0	0	0	0	0
2	0.077	0.374	0.374	0.330	0.330	0.340	0.340	0.325	0.325
3	0.317	0.495	0.121	0.640	0.310	0.640	0.300	0.621	0.296
4	0.467	0.633	0.138	0.679	0.039	0.681	0.041	0.713	0.091
5	0.483	0.604	-0.028	0.692	0.0125	0.665	-0.016	0.722	0.009
6	0.600	0.680	0.076	0.798	0.106	0.779	0.113	0.771	0.048
7	0.617	0.604	-0.076	0.799	0.002	0.777	-0.002	0.800	0.030
8	0.867	0.773	0.169	0.938	0.139	0.976	0.199	0.963	0.163
9	0.900	0.716	-0.057	0.981	0.042	0.917	-0.059	0.945	-0.017
10	1.000	1.000	0.284	1.000	0.0192	1.000	0.083	1.000	0.054
[1] Concentración GPN normalizada [2] Datos Manuales normalizados [3] Análisis Manual [4] Datos GPN-ROI normalizados [5] Análisis GPN-ROI [6] Segmentación <i>K-Means</i> normalizada [7] Análisis <i>K-Means</i> [8] Segmentación Otsu normalizada [9] Análisis Otsu									

Se buscó otra metodología que permitiera identificar si el gel completo puede usarse para detectar la sobre expresión de las proteínas. Para ello, se realizó la segmentación semántica del gel total. Los resultados obtenidos se muestran en la figura 52A, donde se realizaron dos clasificaciones, el ruido de fondo (en color azul) y las proteínas presentes (color amarillo).

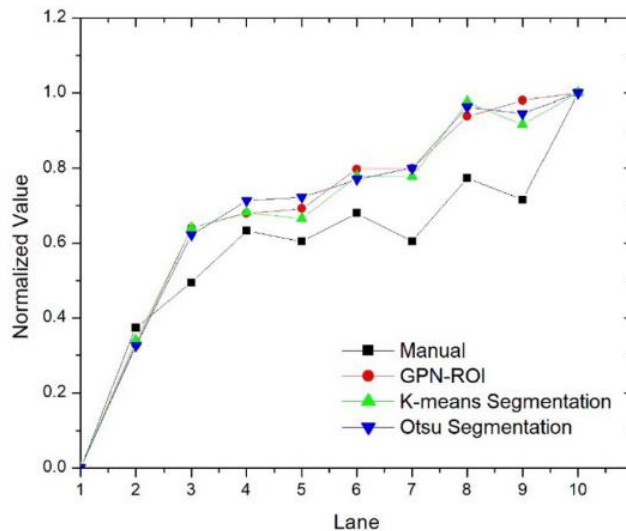


FIGURA 51. GRÁFICA DE DATOS NORMALIZADOS.

Gráfica de la normalización de los datos obtenidos por los métodos manual, GPN-ROI, segmentación de *K-means* y segmentación de Otsu de la tabla 10. Fuente: Elaboración propia.

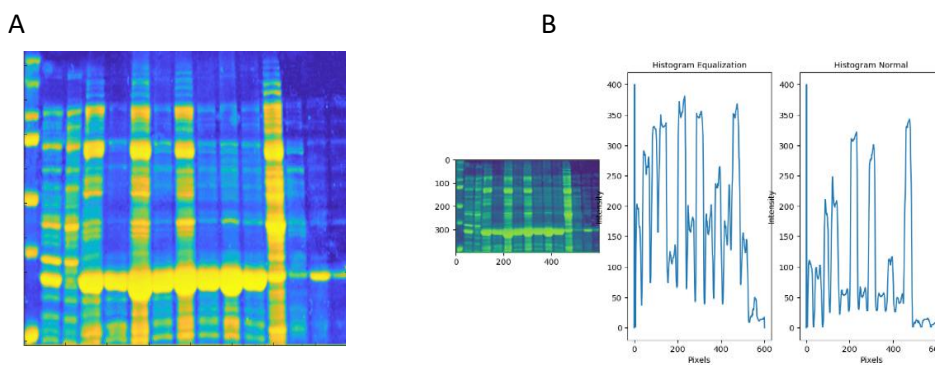


FIGURA 52. SEGMENTACIÓN SEMÁNTICA APLICADA PARA ELIMINAR RUIDO DE FONDO.

A). Segmentación semántica empleada para separar el ruido de fondo de las proteínas. B) Perfil de imagen basado en segmentación de imágenes binarias aplicado a la imagen de la segmentación semántica 52A. Fuente: Elaboración propia.

Los resultados obtenidos siguieron generando múltiples máximos y mínimos (figura 52B), por lo que se le cambió de modelo de color a la imagen obtenida de la segmentación semántica, de RGB a HSV y se realizó una segmentación de color que permitió eliminar totalmente el ruido de fondo como lo muestra la figura 53B.

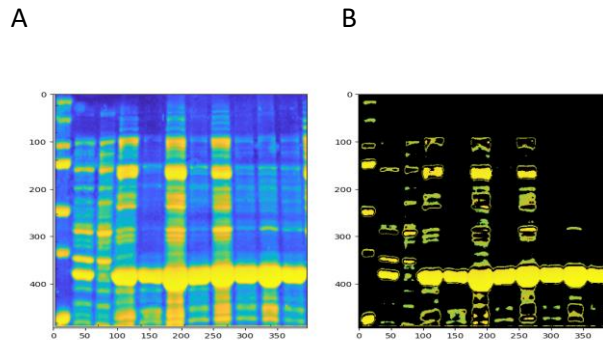


FIGURA 53. ELIMINACIÓN DE RUIDO POR CAMBIO DE MODELO DE COLOR.

A) imagen obtenida empleando la segmentación semántica. B) cambio de modelo de color de RGB a HSV de la imagen 53A para generar un umbral que elimine los colores que no son amarillos. Fuente: Elaboración propia.

A la imagen tratada con HSV se le volvió a aplicar la metodología propuesta y se incrementó el factor estructurante desde un valor de 3 a 25 para intentar delimitar regiones de proteína de un tamaño mayor y eliminar las proteínas detectadas en la parte superior e inferior a la GPN y así intentar detectar su sobre expresión en todas las muestras.

El resultado obtenido se muestra en la figura 54. A valores grandes de factor estructurante al momento de realizar la erosión de la imagen permitió eliminar proteínas de menor tamaño y analizar solamente la proteína GPN que se encuentra expresada a altas concentraciones. Sin embargo, la disminución del ruido conseguida al tratar las imágenes obtenidas por la segmentación semántica también eliminó la proteína de interés cuando la concentración preparada fue menor de 14 mg/ml. Es decir, solamente pudieron detectarse sin ruido las proteínas con una concentración superior o igual a los 14 mg/ml

(ver figura 54A). Esto no impidió detectar la separación entre las muestras preparadas dentro del gel empleando el perfil de imagen basado en segmentación de imágenes binarias (ver figura 54B y 54C). La detección automática de la cantidad de muestras presentes en el gel pudo ser identificada sin necesidad de calcular un umbral sobre la gráfica como fue necesario realizarlo cuando no se empleó la segmentación semántica. Por lo que el uso de la segmentación semántica combinada con el perfil de imagen basado en segmentación de imágenes binarias consiguió la detección automática del total de muestras presentes en el gel. Se almacenaron las posiciones en un arreglo para ser usadas posteriormente para separar cada una de las muestras y hacer un análisis individual de cada proteína GPN sobre expresada y buscar su clasificación tomando como base el orden de sobre expresión de forma más precisa.

Conociendo de forma automática la separación de muestras y el peso molecular de la proteína recombinante analizada, se recortó la región que contiene solamente la proteína GPN expresada a diferentes concentraciones y se erosionó con dos factores estructurantes, un valor pequeño (3, experimentos realizados con la imagen de distintos geles mostraron que este valor fue el adecuado para una buena detección) y un valor grande (25). Los resultados obtenidos se muestran en la figura 55. Al emplear el factor estructurante de 3 y ecualizando la imagen se pudieron detectar las nueve bandas (figura 55A) con áreas muy grandes y un poco de ruido, en cambio, el análisis con un factor estructurante de 25 elimina totalmente el ruido, pero nuevamente no se pudieron detectar las bandas con una concentración menor a 14 mg/ml (solamente se detectaron 7 de 9 bandas).

Conociendo el factor estructurante que presenta mejores resultados y permite detectar todas las bandas de proteína GPN a diferentes concentraciones, se volvió a realizar la segmentación semántica empleando el paquete *Deep Learning* de Matlab R2022a para una región rectangular utilizando 100

imágenes de prueba, 100 etiquetas, 200 de entrenamiento y 400 como etiquetas de entrenamiento para entrenar a la red y poder detectar las clases que corresponden a las bandas y al fondo o ruido. Se obtuvo la segmentación de las proteínas como lo muestra la figura 56A con una exactitud superior al conseguido sin emplear la segmentación semántica para geles heterogéneos (0.95085) como se indica en la tabla 9. Se le aplicó el perfil de imagen basado en segmentación de imágenes binarias (ver gráficas de la figura 56B y 56C) y al conocer tanto el peso molecular de la proteína como la cantidad de muestras presentes en el gel se recortó la región que contiene la proteína GPN a diferentes concentraciones y se analizó cada una de las muestras por separado (ver figura 57) y juntas (figura 56) de la imagen con la proteína GPN a diferentes concentraciones.

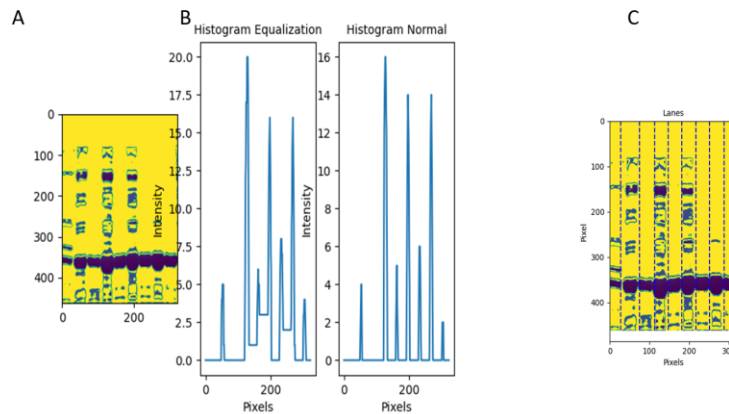


FIGURA 54. CÁLCULO DE NUEVO PERFIL SOBRE LA SEGMENTACIÓN SEMÁNTICA.

A) Aplicación de un elemento estructurante de 25 al momento de realizar la erosión de la imagen obtenida por medio de la segmentación semántica B) Gráficas del perfil de imagen basado en segmentación de imágenes binarias con la imagen 54A ecualizada y sin ecualizar. C) Detección del número de muestras dentro del gel de forma automática sin definir un umbral. Fuente: Elaboración propia.

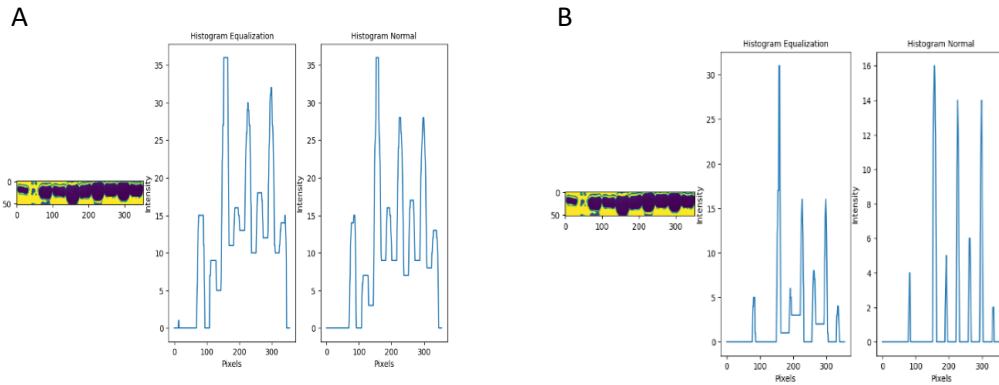


FIGURA 55. CÁLCULO DEL NUEVO PERFIL SOBRE LA REGIÓN ROI OBTENIDA DE LA SEGMENTACIÓN SEMÁNTICA.

A) Factor estructurante de 3 detecta 9 bandas con la imagen ecualizada. B) Factor estructurante de 25 detecta 7 bandas y elimina totalmente el ruido de fondo. Fuente: Elaboración propia.

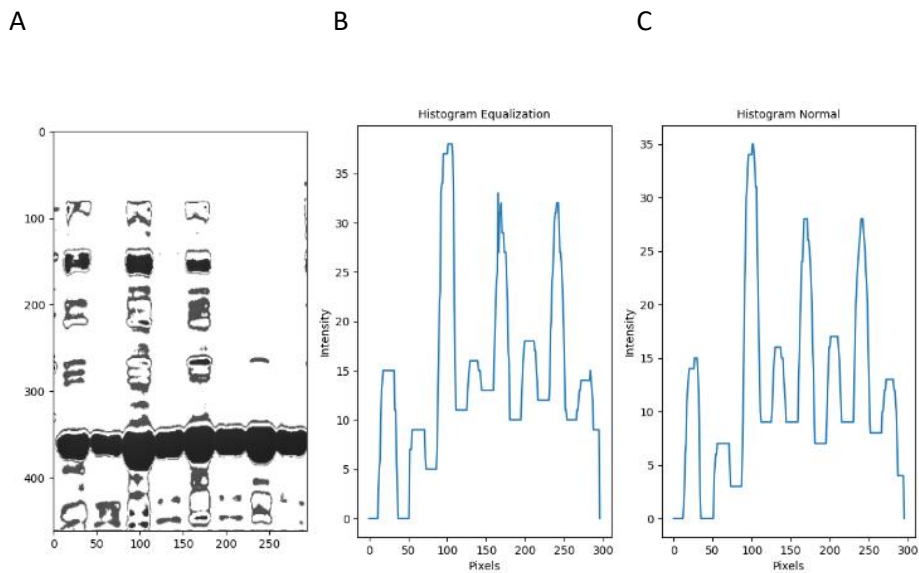


FIGURA 56. APLICACIÓN DEL NUEVO PERFIL SOBRE LA IMAGEN DEL GEL COMPLETO.

A) Imagen del gel SDS-PAGE con la proteína GPN a diferentes concentraciones tratada por segmentación semántica y por el perfil de imagen basado en segmentación de imágenes binarias. B) Gráfica del PIBSIB obtenida con la imagen ecualizada. C) Gráfica PIBSIB obtenida con la imagen sin ecualizar. Fuente: Elaboración propia.

A cada una de las diferentes muestras que incluyen la proteína GPN a diferentes concentraciones se les aplicó el perfil de imagen basado en segmentación de imágenes binarias para identificar mediante el valor máximo su sobre expresión. Para encontrar un solo máximo absoluto se tomaron todos los valores y se promediaron, lo que permitió detectar un solo máximo local y se comparó este valor con todos los demás tomando el factor estructurante de 3 (previamente se había demostrado que es el valor que permitió detectar todas las muestras).

Los resultados de los máximos obtenidos (ver figura 57) permitieron identificar la sobre expresión de la proteína recombinante de forma correcta, los valores conseguidos denominados *semantic segmentation* (ROI with GPN) fueron 11.325 (2 mg/ml), 22.5 (9.5 mg/ml), 27.15 (14 mg/ml), 27.35 (14.5 mg/ml), 30.275 (18 mg/ml), 30.675 (18.5 mg/ml), 36.15 (26 mg/ml), 36.8 (27mg/ml), 38.55 (30 mg/ml). Por lo tanto, el emplear la segmentación semántica seguida de la metodología desarrollada permitió separar cada una de las muestras e identificar de forma correcta y sin errores la sobre expresión de la proteína GPN.

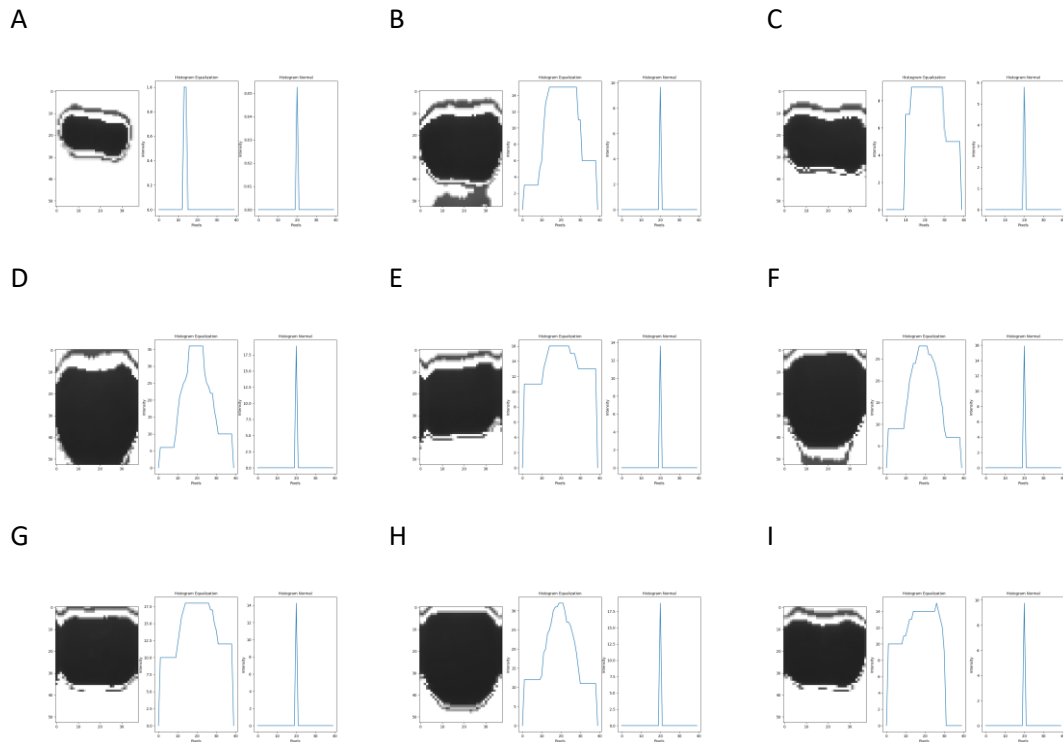


FIGURA 57. CÁLCULO DEL NUEVO PERFIL SOBRE LAS MUESTRAS DE FORMA INDEPENDIENTE. Perfil de imagen basado en segmentación de imágenes binarias aplicada al análisis individual de la imagen obtenida mediante la técnica de segmentación semántica (figura 56A) para encontrar la sobre expresión de la proteína GPN2 a diferentes concentraciones. Fuente: Elaboración propia.

El éxito de los resultados obtenidos indica que el método desarrollado (perfil de imagen basado en segmentación de imágenes binarias) permite identificar tanto la posición de la proteína de interés como su sobre expresión, por lo que el método desarrollado puede ser empleado como diagnóstico para identificar si una proteína relacionada con cáncer de mama se está sobre expresando lo que indicaría que se está agravando el cáncer llegando a metástasis o se encuentra cambiando de estadio.

Esta metodología también puede emplearse para identificar si un tratamiento anti cancerígeno está funcionando, ya que se esperaría que la sobre expresión

en las muestras vaya disminuyendo o que mantenga valores similares en los máximos detectados indicando que no está aumentando la metástasis o que está disminuyendo el incremento de células cancerígenas siempre y cuando el tipo de cáncer de mama estudiado se encuentre relacionado con alguna proteína específica y que su banda pueda identificarse en un gel de poliacrilamida.

Estos resultados indican que si la metodología planteada permite identificar que una banda dentro de un gel se está incrementando (sobre expresándose) entonces puede ser empleada en imágenes médicas relacionadas con diferentes tipos de cánceres donde el número de células vayan incrementando por causa de la misma enfermedad. Para ello, se analizaron las siguientes imágenes:

- a) Se tomó la imagen de la figura 5 de [68] recuperada de [69] donde muestran que el empleo de *hypoxia methyl sulfone* reduce los niveles de una proteína asociada con el incremento de metástasis de cáncer de mama. La tabla 11 muestra las células control y las células que incluyen *methyl sulfone* (que impide el desarrollo de la metástasis) antes y después. Los resultados experimentales de Caron & Caron en el 2015 mostraron que la metástasis se detiene y en caso de no emplear la *methyl sulfone* la metástasis se incrementa (ver tabla 12). En las imágenes de [68] se implementó nuestra metodología que presentó los resultados mostrados en las tablas 11 y 12.

TABLA 11. VALORES OBTENIDOS EN CUATRO IMÁGENES DE CÉLULAS QUE NO HAN ENTRADO EN METÁSTASIS. FUENTE: [68]

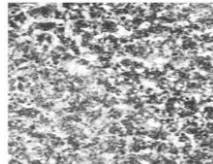
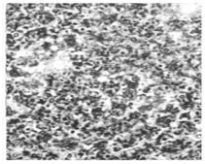

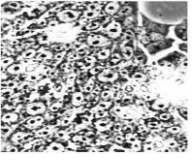
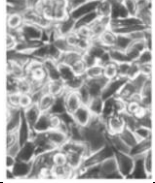
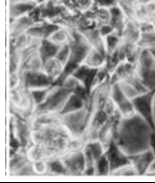
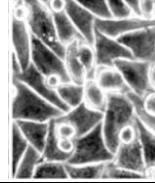
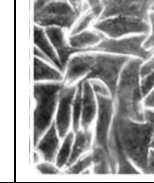
Muestra	Control	Control	<i>Methyl sulfone</i>	<i>Methyl sulfone</i>
Datos	101.9585	97.578514	69.24051	69.778725
Imagen				

TABLA 12. VALORES OBTENIDOS EN CUATRO IMÁGENES DE CÉLULAS QUE PADECEN METÁSTASIS. FUENTE: [68]

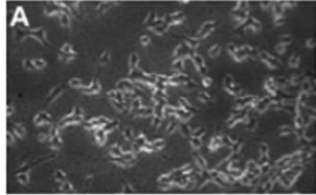
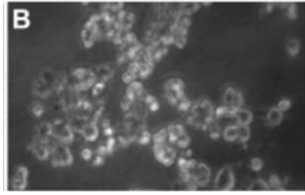
Muestra	Control	Control	<i>Methyl sulfone</i>	<i>Methyl sulfone</i>
Datos	179.51479	181.96286	200.70326	201.60355
Imagen				

Nuestros resultados de las imágenes analizadas coincidieron con los resultados experimentales obtenidos por [68], donde emplearon cultivos celulares, inmunoblot, análisis micro ARN 210 buscando la expresión de genes y estudios de inmunofluorescencia para verificar que las células presenten metástasis o no; y nuestra metodología solamente midió los máximos de los picos promediados de las imágenes para detectar cuales células no sufren de cáncer (tabla 11) y aquellas que si se afectaron, el valor de los máximos fue incrementando en las células tal y como fue aumentando la metástasis (tabla 12).

- b) Se aplicó nuestra metodología en las imágenes obtenidas en la figura 2 de [70] en melanoma de piel humana con líneas celulares A375 (ver tabla 13). Sus resultados mostraron que cuando las células no son tratadas con nanopartículas no entran en apoptosis y la enfermedad se mantiene o se agrava.

Nuestro análisis encontró un máximo mayor en la imagen que incluye las células dañadas y un menor valor en las células normales por lo que puede emplearse para detectar daño celular.

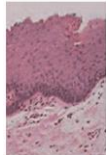
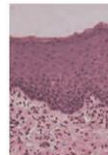
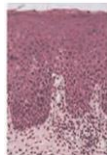
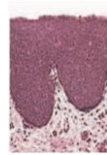
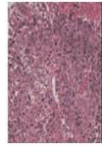
TABLA 13. IMÁGENES DE MELANOMA DE PIEL HUMANA LÍNEA CELULAR A375 Y DATOS OBTENIDOS. FUENTE: [70]

Muestra	Células normales	Células con membrana celular dañada
Promedio de máximos de color blanco resultado de aplicar el perfil de imagen basado en segmentación de imágenes binarias	167.21768	180.85349
Imagen		

- c) Finalmente, se emplearon las imágenes de la figura 2 de [71] que representan la progresión del carcinoma de células escamosas orales de hiperplasia epitelial a través de diferentes etapas de displasia. Sus resultados mostraron como se incrementó el daño celular a través de los siguientes estados: Hiperplasia, displasia leve, displasia moderada, displasia severa e invasiva.

Se aplicó el perfil de imágenes basado en segmentación de imágenes binarias a cada una de las imágenes y se obtuvieron los datos mostrados en la tabla 14.






TABLA 14. PROGRESIÓN DEL CARCINOMA DE CÉLULAS ESCAMOSAS ORALES DE HIPERPLASIA EPITELIAL A TRAVÉS DE DIFERENTES ETAPAS DE DISPLASIA. FUENTE [71].

Muestra	Hiperplasia	Displasia leve	Displasia moderada	Displasia severa	Invasiva
Datos	132	144	161	178	205
Imagen					

Nuestros datos corroboraron el incremento en el daño celular, los valores obtenidos al aplicar la metodología propuesta pudieron detectar el incremento en el número de células que ocurre cuando va aumentando el estado cancerígeno.

Se tomaron cortes al azar con un tamaño de 23x35 pixeles de cada uno de los diferentes estados de displasia y se volvió a aplicar el perfil de imagen basado en segmentación de imágenes binarias para identificar si en regiones pequeñas puede detectarse el incremento en el daño celular, los resultados obtenidos se muestran en la tabla 15.

TABLA 15. DATOS DE LOS CORTES ELEGIDOS AL AZAR DE LOS DIFERENTES ESTADOS DE DISPLASIA. FUENTE: [71]

Muestra	Hiperplasia	Displasia leve	Displasia moderada	Displasia severa	Invasiva
Datos	8.826087	13.56521	13.869565	19.913044	17.95652
Imagen					

Al analizar estas regiones elegidas al azar, los datos (obtenidos al aplicar el PIBSIB), pudieron detectar el incremento en el daño celular ocurrido en los estados de hiperplasia (8.826087), displasia leve (13.56521), displasia moderada (13.869565) y displasia severa (19.913044) sin problema, ya que los valores de los máximos obtenidos en cada imagen fueron incrementando tal y como fue reportado el daño celular. Sin embargo, el último estado (invasivo) el PIBSIB lo reportó entre moderado y severo, esto puede explicarse porque la imagen original (antes de extraer una región) es la única que no contiene dos diferentes tipos de líneas celulares, lo que imposibilitó a la metodología propuesta poder detectar el incremento en el daño celular.

Capítulo 5. Discusión, conclusiones y trabajo a futuro

5.1. Discusión

Se desarrolló una metodología que permitió construir un buffer para conseguir la purificación de la proteína recombinante GPN expresada en la bacteria *Escherichia coli*, el método se basa en el efecto que tienen los diferentes aditivos que se han utilizado y su influencia sobre los enlaces y/o solubilidad que presentan tanto los contaminantes como la proteína misma.

El primer factor que se analizó fue la temperatura de crecimiento para conseguir la proteína de interés de forma soluble. Los resultados indicaron que para obtener la GPN dentro de la fracción soluble fue necesario realizar el crecimiento a 10°C después de haber inducido la expresión de la proteína recombinante con IPTG.

El segundo punto por tomar en cuenta fue medir la solubilidad de la proteína de interés dentro del buffer inicial propuesto a base de Tris-HCl en una concentración de 100 mM para mantener el pH a 8.2. Con las diferentes concentraciones empleadas de NaCl para incrementar la solubilidad de todo el extracto celular se encontró que a una concentración de 100 mM de NaCl se consigue la mayor cantidad de proteínas en su forma soluble y que a valores mayores empieza a precipitar.

Teniendo el valor de la concentración salina que proporcionó una mayor fracción soluble, se realizaron variaciones en el mismo buffer Tris-HCl para identificar su efecto sobre la eliminación de contaminantes, a partir de 100 mM de Tris-HCl no se logró disminución de impurezas por lo que hasta el momento el buffer para los siguientes experimentos se trabajó a 100 mM Tris-HCl y 100 mM NaCl que es el que permitió mantener la menor cantidad de contaminantes y la mayor cantidad de GPN soluble.

Posteriormente, se usaron diferentes concentraciones de EDTA para reducir el daño de oxidación, inhibir las metaloproteasas evitando que se desnaturalice la proteína estudiada y medir su efecto en la eliminación de impurezas. Estos resultados mostraron que para la GPN el empleo de 400 μM de EDTA ayudó en la reducción del grosor de algunas de las bandas de proteínas consideradas contaminantes manteniendo la misma concentración de la GPN. Las concentraciones empleadas fueron reducidas ya que en cantidades grandes libera el níquel de la columna impidiendo realizar la cromatografía IMAC.

Al seguirse presentando bandas del mismo grosor que la GPN y no poder ser eliminadas, se concluyó que posiblemente se encontraran enlazadas mediante enlace disulfuro sobre la proteína GPN, por lo que para separarlas se emplearon diferentes concentraciones de DTT. Este reactivo permite mantener el estado activo de la proteína recombinante, eliminar enlaces disulfuros correspondientes a proteínas contaminantes unidas sobre la proteína de interés y medir su efecto en la eliminación de estos; los valores empleados fueron pequeños para evitar la desnaturalización de la GPN. Los resultados indicaron que el empleo de 100 μM favorecen la eliminación de las proteínas contaminantes pudiendo conseguir la proteína GPN de forma pura a una concentración de 2mg/ml.

Para mejorar la solubilidad de la proteína, se buscó el efecto que tienen diferentes aminoácidos sobre el proceso de purificación de la proteína recombinante, el empleo de Glutamato de Sodio tuvo como propósito incrementar la fuerza iónica de la muestra, estabilizar la proteína recombinante y disminuir su producción dentro de cuerpos de inclusión. La arginina se empleó por sus propiedades similares al Glutamato. Sin embargo, el estudio realizado con arginina indicó que no reduce contaminantes de forma significativa y en cambio disminuye la cantidad de GPN dentro de la muestra y la elimina. Un efecto similar en la eliminación de contaminantes se consiguió

empleando 100 mM de Glutamato de Sodio. Para buscar un reemplazo de la arginina que mantuviera la misma carga, pero con menor fuerza, se utilizó el aminoácido lisina ya que es menos polar que la arginina. Al emplear este aminoácido a una concentración de 100 mM se consiguió disminuir la cantidad de impurezas y conservar una mayor cantidad de la proteína recombinante que al emplear la arginina.

Cuando se realizó la purificación empleando solamente 100 mM Tris-HCl, 100 mM NaCl, 100 mM Lisina y 100 mM Glutamato de Sodio a un pH de 8.2 se consiguió eliminar una cantidad considerable de contaminantes manteniendo dos bandas de proteínas que se siguieron conservando y que al realizar la cinética empleando DTT pudieron eliminarse. Esto sugiere que estas proteínas se encontraban unidas mediante enlaces disulfuro sobre la proteína GPN y que por lo tanto pudieron ser eliminadas utilizando DTT o 2-Mercaptoethanol lo cual depende de la resina que se esté utilizando al realizar la cromatografía de afinidad de metales inmovilizados.

Con todas las mediciones realizadas se determinó que el buffer ideal para conseguir la proteína GPN de forma pura y a altas concentraciones deberá tener los siguientes reactivos: 100 mM Tris-HCl, glicerol al 5%, Tritón X-100 a 0.2%, 100 mM NaCl, 400 μ M EDTA, 100 μ M DTT a un pH de 8.2. Se agregó el glicerol para darle estabilidad al buffer y evitar que se aglomeren las proteínas. El Tritón X-100 para prevenir la precipitación, solubilizar proteínas insolubles y liberar proteínas presentes dentro de cuerpos de inclusión.

Los resultados muestran que el empleo de este buffer creado para la proteína GPN recombinante expresada en *Escherichia coli* consiguió, sin empleo de tabletas inhibitoras de proteasas, la proteína de forma totalmente pura y homogénea, con baja polidispersión corroborada por análisis realizados de dispersión dinámica logrando una concentración de 30 mg/ml por lo que puede utilizarse para estudios de cristalografía o de genómica estructural.

La proteína pura se concentró a diferentes volúmenes para emular muestras de diferentes etapas de cáncer de mama tipo Her2+ que resulta altamente invasivo catalogado de tipo carcinoma ductal y lobular invasivo (CDI y CLI). Se realizó la electroforesis de las distintas muestras y se prepararon los geles SDS-PAGE para realizar el estudio de análisis de imágenes.

Se prepararon 382 geles con 10 muestras en cada uno de ellos, haciendo un total de 3,820 muestras, de las cuales 2,259 salieron defectuosas, decoloradas o con poca expresión de la proteína GPN. 1,561 muestras presentaron la proteína GPN expresada o sobre expresada, pero con diferentes tonalidades de color, por lo que se denominaron geles heterogéneos. De esta muestra de geles se tomaron 669 muestras que presentaron mismas condiciones de color y se denominaron geles homogéneos.

Se realizaron concentraciones controladas de la proteína GPN con los valores de 0, 2, 9.5, 14, 14.5, 18, 18.5, 26, 27, 30 mg/ml de solución. Las concentraciones se agregaron al extracto bacteriano para conseguir muestras que representaron diferente sobre expresión de la proteína recombinante y fuera representativo de distintos estadios de cáncer de tipo Her2+.

Los perfiles de intensidad calculados en las imágenes no permitieron obtener características que pudieran identificar ya fuera el número de muestras por gel o el número de bandas de proteína por muestra, debido a la presencia de proteínas endógenas dentro de la misma muestra. Sólo en el caso de que no hubiera muestras con contaminantes (proteínas endógenas), sin ruido de fondo (el mismo color sin objetos), y con proteínas bien definidas en la muestra (como en la muestra control) es como con el perfil de la imagen se pudo encontrar la posición de las bandas de proteínas. Sin embargo, este no es el caso cuando se obtienen muestras de pacientes, ya que siempre existen las proteínas que requiere la célula del tejido que se encuentre analizando para poder realizar sus funciones metabólicas.

Hasta el momento no se encontró la descripción del uso de herramientas computacionales para realizar análisis del número de muestras por gel o alguna banda de proteína en específico, en las imágenes de geles de proteínas de forma automática. Además, ninguno de los métodos desarrollados permite encontrar la sobre expresión de alguna proteína de interés biológico. Por lo anterior, se decidió desarrollar una metodología para realizar el análisis de imágenes de los geles de proteínas iniciando con un preprocesamiento de la imagen. Se invirtieron los colores de las imágenes en escala de grises y fueron binarizadas para lograr una mejor separación de las muestras presentes en el gel empleando técnicas de erosión y dilatación. Después, tomando en cuenta que en la separación de las muestras o bandas de proteínas en la imagen del gel debe haber un aumento en el color negro o, lo que es lo mismo, una disminución en el color blanco, se desarrolló una técnica que permitió encontrar estas variaciones en el color blanco (relacionadas con la separación de las muestras o de las bandas de proteínas) denominada “Perfil de imagen basado en segmentación de imágenes binarizadas, PIBSIB”.

La gráfica generada con la metodología desarrollada permitió relacionar los múltiples mínimos con la cantidad de muestras presentes en el gel. Al elegir alguna de las muestras y repetir la metodología desarrollada en ella, la gráfica que representó al nuevo perfil permitió relacionar ahora los múltiples máximos con la posición de cada una de las bandas de proteínas presentes en la imagen del gel.

El nuevo perfil de imagen no solamente detectó la cantidad de muestras presentes en el gel y la cantidad de bandas de proteínas, si no que permitió relacionar el tamaño de cada uno de los máximos con la cantidad de proteína expresada en la muestra, es decir, a mayor tamaño mayor concentración. Se comprobó su funcionamiento con una segunda proteína, la BSA, donde el

tamaño de los máximos locales en la gráfica correspondió a la cantidad de proteína expresada.

El perfil de imagen basado en segmentación de imágenes binarias, PIBSIB, permitió encontrar la cantidad sobre expresada de la proteína GPN cuando se presentaron en el gel las diferentes concentraciones de la proteína. Sin embargo, la base de los máximos locales no inició en el origen y se tuvo que elegir un umbral que permitió eliminar los múltiples máximos. Este umbral hizo que se detectará de forma automática la cantidad de muestras presentes en el gel.

Conociendo la posición del número de muestras que tiene el gel, se eligió la primera muestra que contiene al control del peso molecular de las proteínas y se empleó la metodología propuesta para detectar las proteínas del control, conocer el peso molecular de las mismas y desarrollar un método de interpolación que permitió conocer el tamaño de las proteínas presentes en el resto de las muestras.

Habiendo identificado el peso molecular de la proteína GPN, se eligió la región que contiene todas las muestras que la incluyen a diferente sobre expresión, se aplicó el perfil de imagen basado en segmentación de imágenes binarias y pudieron relacionarse el tamaño de los múltiples máximos con la cantidad de proteína expresada por muestra. Esto demostró que la metodología permitió detectar la sobre expresión de la proteína de interés de manera correcta siempre y cuando no exista ruido de fondo. Si esta condición se cumple entonces se puede detectar el nivel de sobre expresión junto con el orden de expresión.

Se evaluó la exactitud del método por medio de la matriz de confusión empleando geles homogéneos, es decir, que presentaron el mismo color, sin distorsiones y se consiguió un valor de 0.985052. Se repitió el análisis

empleando el total de muestras incluyendo los geles heterogéneos y el valor de la exactitud bajó a 0.91736.

Se comparó la metodología propuesta con otros métodos que incluyeron el cálculo de áreas de las muestras de la proteína de forma manual, empleando segmentación de *K-means* y segmentación de Otsu. Sin embargo, el perfil de imagen basado en segmentación de imágenes binarias proporcionó mejores resultados encontrando el orden de sobre expresión de la proteína GPN que los métodos restantes no consiguieron. Además, fue necesario para cada uno de los métodos utilizados haber recortado cada una de las imágenes que contienen la proteína GPN con diferentes concentraciones ya que no existe metodología que permita analizar dichas muestras dentro de un gel SDS-PAGE de proteínas.

Para mejorar la metodología propuesta, se aplicó segmentación semántica para eliminar el ruido de fondo y detectar solamente las proteínas presentes en el gel, se complementó transformando las imágenes obtenidas del espacio de color RGB a HSV y se realizó un filtro de color que permitió eliminar por completo las proteínas no deseadas en cada una de las muestras. En estas imágenes se aplicó nuevamente el perfil de imagen basado en segmentación de imágenes binarias consiguiendo una nueva gráfica libre de ruido donde el tamaño de los máximos locales se relacionó con la cantidad de proteína expresada dentro de cada muestra y se detectó el orden de sobre expresión de la proteína.

Después de corroborar que el perfil de imagen basado en segmentación de imágenes binarias permitió identificar el grosor o tamaño de las bandas, se planteó la posibilidad de utilizarla para identificar la presencia de grandes cantidades celulares relacionadas con diferentes tipos de cánceres, ya que este padecimiento se relaciona también con el incremento del número de células afectadas, se verificó si la metodología desarrollada tiene la capacidad de detectar dicho incremento celular. Este planteamiento surge debido a que

al incrementar el número de células y tener la imagen binaria y con colores invertidos, los espacios existentes entre ellas deben de ir disminuyendo por estar incrementando el número celular, esto estaría relacionado con el incremento en el color blanco dentro de la imagen.

Se tomaron imágenes de muestras celulares donde se analiza de forma visual la cantidad de células obtenidas al momento de aplicar una solución que inhibe el crecimiento de células cancerígenas de mama. El perfil de imagen basado en segmentación de imágenes binarias pudo detectar la disminución en el número de células al relacionar la imagen con el tamaño del pico en la gráfica. Los valores encontrados estuvieron en concordancia con el estado en que se encontraron las células, es decir, en muestras donde no se tienen células afectadas los valores del tamaño de los máximos fue menor que en los casos donde se tienen las células cancerígenas.

Los mismos resultados se encontraron al analizar diferentes tipos de muestras celulares. Por tal motivo, el perfil de imagen basado en segmentación de imágenes binarias permitió no sólo detectar la sobre expresión de proteínas en geles SDS-PAGE, sino que también tuvo la capacidad de detectar el incremento en el número de células dentro de la imagen.

Los programas desarrollados Scanalytics, GelcomparII, Gel-Pro Analyzer, TotalLab, PDQuest, Proteomweaver, Dcyder 2D, imageMaster, Melanie, BioNumerics, Redfin, Gel IQ, Z3 y Delta2D Flicker están dirigidos al análisis de imágenes de geles de DNA, en la identificación de bandas en geles de proteínas se aplican distintos filtros para eliminar el mayor ruido posible. Además, todos son semiautomáticos, necesitan que un operador interactúe con el programa al elegir alguna columna o banda de interés. El usuario también ajusta los cambios de intensidad y decide el mejor umbral que contiene menor ruido de fondo. Además, algunos de los programas realizados son muy complicados para los analistas que no tienen conocimiento sobre

intensidades y umbralización, en consecuencia, no consiguen un buen análisis del gel [33, 29].

Por lo tanto, retomando la tabla comparativa y agregando la metodología planteada queda como se muestra en la tabla 16.

TABLA 16. TABLA COMPARATIVA COMPLETA.

Programas y avances de los trabajos desarrollados en el área de análisis de imágenes de geles de ADN y proteínas. Fuente: Elaboración propia.

Autor/SW	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
PyElph, 2012	N	N	N	N	N	N	N	N	N	-	-	-	-
GelCluster, 2013	S	S	N	N	N	N	N	N	N	-	-	-	-
Sim, 2015	N	S	N	N	N	N	N	N	N	N	S	S	S
Abadi, 2015	S	S	-	S	N	N	N	N	N	N	S	S	S
Intarapanich, 2015	S	S	S	N	N	N	N	N	S	N	S	N	S
Turan, 2016	N	S	N	N	N	N	N	N	N	N	S	N	S
Fernández-Lozano, 2016	N	N	N	N	N	N	N	N	S	N	N	S	N
Rezaei, 2016	S	N	N	N	N	N	N	N	S	N	S	N	S
Verbeek, 2017	N	S	-	N	S	N	N	N	N	N	S	S	N
Goez, 2018	N	N	S	N	N	N	N	N	N	N	N	S	N
Ou, 2020	N	S	S	N	N	N	N	N	N	N	N	S	N
Alnamoly (EGBiolmage), 2020	S	S	S	S	S	N	N	N	N	S	S	N	S
GelAnalyzer, 2022	S	N	N	S	N	N	N	N	N	-	-	-	-
JJuárez, 2023	S	S	N	S	N	S	S	S	S	S	S	S	N

C1: Detección de columnas
 C2: Detección de bandas
 C3: Requiere el usuario colocar el nombre de las muestras
 C4: Proporciona el peso molecular de forma automática
 C5: Realiza la agrupación de las bandas empleando K-means
 C6: Usa el buffer de carga para calcular el peso molecular
 C7: Usa la imagen del gel para encontrar las proteínas de baja expresión
 C8: Usa la imagen del gel para encontrar proteínas sobre expresadas
 C9: Separa regiones concentradas de proteínas
 C10: Emplea métodos de interpolación para encontrar el peso molecular
 C11: Es utilizado para analizar geles de una dimensión
 C12: Es utilizado para analizar geles de proteínas
 C13: Es utilizado para analizar geles de ADN
 S: Sí
 N: No

De los programas desarrollados, los más completos son los de Alnamoly, GelAnalyzer y JJuárez. De estos, solamente Alnamoly y la propuesta que se

hace en este trabajo son capaces de detectar tanto las columnas como las bandas del gel. Sin embargo, la propuesta de esta investigación doctoral no requiere que el usuario coloque el nombre de las muestras superando a Alnamoly, además no requiere de agrupar las bandas para poder detectarlas, puede encontrar la sobre expresión de proteínas lo que no realiza ninguno de los anteriores y hasta el momento solamente ha sido empleado para geles de proteínas y no de ADN.

Así, finalmente, el perfil de imagen basado en la segmentación de imágenes binarias tiene la capacidad de detectar el nivel de sobre expresión de bandas de proteína dentro de un gel SDS-PAGE, puede identificar el número de muestras y el número de bandas de proteínas. Además, puede identificar si una muestra obtenida de algún tejido que contenga un padecimiento relacionado con el cáncer está empeorando o mejorando, por lo que puede emplearse como método de diagnóstico confiable, ya que si las muestras son analizadas por laboratoristas de forma manual se pueden confundir y no detectar si la enfermedad está avanzando o si el paciente está presentando alguna mejora.

5.2. Conclusiones

Se construyeron muestras que emulan la sobre expresión de proteínas en geles de poliacrilamida o SDS-PAGE para representar diferentes concentraciones de una proteína de interés biológico. Para conseguirlas se utilizó la proteína GPN que se encuentra relacionada con el cáncer de mama tipo Her2+ considerado altamente peligroso e invasivo, tanto del conducto como de los lóbulos mamarios (Carcinoma ductal invasivo y carcinoma lobular invasivo).

Para conseguir la proteína GPN en forma soluble, pura, homogénea y monodispersa se analizaron las fuerzas presentes en los enlaces químicos para crear un buffer que permitió mantener activa la proteína y libre de contaminantes. El buffer final obtenido incluyó los siguientes reactivos: 100 mM Tris-HCl, glicerol al 5%, Tritón X-100 a 0.2%, 100 mM NaCl, 400 μ M EDTA, 100 μ M DTT a un pH de 8.2. La metodología desarrollada se considero de alta importancia en consecuencia se sometio al proceso de patentamiento. El proceso continua, pasando ya el examen de forma.

Con las muestras se crearon geles a diferentes concentraciones que emularon los distintos estadios del cáncer de mama relacionado con la sobre expresión de la proteína y se realizó el análisis de imágenes de los geles.

Se desarrolló una metodología denominada perfil de imagen basado en segmentación de imágenes binarias, que consistió en emplear una máscara binaria de 1x400 pixeles que fue recorriendo la imagen píxel por píxel hasta analizarla toda. En cada recorrido binarizó y erosionó la región analizada para generar un arreglo de datos y elegir la intensidad del color blanco del último valor del arreglo. Estos valores fueron graficados para generar un nuevo perfil de la imagen donde los múltiples valores mínimos representaron la separación de todas las muestras presentes en el gel. La misma metodología se aplicó a cada muestra para detectar la posición de las proteínas o bandas (cambiando el tamaño de la máscara binaria a 1x50 pixeles) y encontrar la proteína de interés con base en su peso molecular (con un error de 3.35%). De forma automática se pudo encontrar la posición de la proteína y tomando los máximos del perfil de imagen basado en segmentación de imágenes binarias se pudo encontrar la sobre expresión de la proteína GPN y ordenarlas en base a la concentración que presentaron en cada muestra.

Cuando los geles fueron homogéneos se consiguió una exactitud de 0.985052 medida mediante una matriz de confusión. Si los geles fueron heterogéneos e incluyeron errores de color o diseño la exactitud fue de 0.91736, este valor fue

mejorado en los geles solamente cuando se aplicó la segmentación semántica obteniendo el valor de 0.95085.

La técnica desarrollada permitió encontrar la proteína deseada con base en su peso molecular en el gel y verificar su nivel de sobre expresión. Se comparó con el cálculo manual de áreas, por segmentación *K-means*, segmentación de Otsu y segmentación semántica resultando el perfil de imagen basado en segmentación de imágenes binarias más eficiente para encontrar la sobre expresión de la proteína de interés en geles homogéneos; para los geles heterogéneos fue ligeramente superada la metodología propuesta al utilizar la segmentación semántica analizando las proteínas de todas las muestras por separado y promediando los valores obtenidos para conseguir un solo máximo.

Por lo tanto, la técnica desarrollada es capaz de identificar cuál muestra contiene una mayor concentración de proteína, por lo que puede emplearse como un nuevo método de diagnóstico para identificar si un paciente se encuentra en una etapa agresiva del cáncer de mama tipo Her2+. Incluso puede emplearse para identificar si el empleo de alguna terapia o medicamento está funcionando para mejorar la salud del paciente, ya que, si la proteína GPN se encuentra disminuyendo en su concentración, es sinónimo de que el paciente tiene mejoras sobre la enfermedad.

Se aplicó la metodología para comparar imágenes que muestran células de mama normales y células cancerígenas pudiendo detectar como se va incrementando la metástasis. También se probó con imágenes de melanoma de piel humana en líneas celulares A375 donde pudo comprobarse el incremento en el daño celular.

Por último, el perfil de imagen basado en segmentación de imágenes binarias pudo identificar la progresión del carcinoma en diferentes niveles de

hiperplasia epitelial analizando la imagen completa y cortando pequeñas regiones de 25x35 píxeles.

Todos estos resultados indican que la metodología propuesta puede ser empleada para identificar la sobre expresión de proteínas de interés biológico y con ello detectar en muestras de pacientes con cáncer si la enfermedad va avanzando o incluso si algún tratamiento está funcionando siempre y cuando el tipo de cáncer se encuentre relacionado con la sobre expresión de alguna proteína específica. El perfil de imagen basado en segmentación de imágenes binarias también puede ser empleado para identificar en imágenes de tejidos celulares si la cantidad de células está aumentando lo que es señal de que el cáncer está incrementando.

Los resultados obtenidos indican que las metodologías desarrolladas pueden ser empleadas como método de diagnóstico para encontrar la sobre expresión de alguna proteína de interés biológico relacionada con algún padecimiento de cáncer y como puede ser también empleada para detectar en imágenes de células los diferentes estadios de diferentes tipos de cáncer, se han superado los objetivos planteados creados a partir de la pregunta de investigación.

Por lo tanto, después del análisis realizado, se alcanzaron los objetivos de la tesis respondiendo positivamente a la hipótesis de investigación, así como a la pregunta de investigación consiguiendo un sistema que puede ser empleado como método de diagnóstico para identificar la sobre expresión de proteínas relacionadas con algún cáncer de mama. Esta técnica se puede usar no solamente para detectar la proteína GPN o alguna otra GTPasa, sino que puede ser empleada para detectar la sobre expresión de otro tipo de proteínas relacionadas con diferentes tipos de cánceres como puede ser el cáncer de próstata. Adicionalmente, esta metodología resulta ser más económica para detectar proteínas consideradas biomarcadores de enfermedades que al buscarlas empleando técnicas moleculares.

5.3. Trabajo futuro

Después de la investigación desarrollada y como el análisis de imágenes de geles de proteína SDS-PAGE es todavía un terreno fértil, se pueden desprender las siguientes líneas de investigación para ser realizadas como una extensión de la presente investigación:

1. Realizar una corrección de geles dañados por mala tinción o por efecto de cara feliz.
2. Identificar el total de proteínas presentes en una muestra.
3. Construir bases de datos de proteínas detectadas en una muestra y relacionarlas con las proteínas identificadas con distintos tipos de cáncer.
4. Separar proteínas con peso molecular similar que ocupen regiones cercanas.
5. Aplicar diferentes redes neuronales empleando segmentación semántica y comparar su eficiencia.
6. Emplear segmentación semántica para separar proteínas de interés en la muestra de las proteínas no deseadas
7. Emplear el perfil de imagen basado en segmentación de imágenes binarias, desarrollado en esta tesis, en muestras de tejidos celulares para detección de distintos niveles de cáncer.

Apéndices

A. Lista de Figuras

Figura 1. Geles de electroforesis de genes.	10
Figura 2. Incidencias y defunciones de cáncer de mama a nivel internacional.	11
Figura 3. Flujo de incidencia y mortalidad del cáncer de mama a nivel internacional.....	12
Figura 4. Comparación entre células normales y afectadas.	13
Figura 5. Supervivencia relacionada con sobre expresión de proteína GPN.	14
Figura 6. Geles SDS-PAGE dañados.	15
Figura 7. Diferentes tipos de geles.	26
Figura 8. Bacteria modificada.	27
Figura 9. Enlace proteína-columna.	28
Figura 10. Columna IMAC.	28
Figura 11. Segmentación semántica aplicada a la detección de objetos.	38
Figura 12. Segmentación semántica aplicada en biomedicina.	38
Figura 13. Segmentación semántica separando objetos de misma clase. ...	39
Figura 14. Representación de una CNN aplicada a la segmentación semántica.	40
Figura 15. Arquitectura U-net empleada para el análisis de imágenes médicas.	41
Figura 16. Diagrama de flujo de la primera parte de la metodología.	44
Figura 17. Diagrama de flujo del procedimiento para conseguir la muestra final.	47
Figura 18. Diagrama para obtener muestras en geles SDS-PAGE.	52
Figura 19. Diagrama de flujo de la segunda parte de la metodología.	54
Figura 20. Esquema general. Fuente: Elaboración propia.	56

Figura 21. . Algoritmo general.....	58
Figura 22. Esquema general de red neuronal.....	59
Figura 23 Prueba de expresión de proteína GPN.	61
Figura 24. Prueba de expresión de la proteína GPN a diferentes temperaturas.....	62
Figura 25. Análisis con diferentes concentraciones de NaCl.	63
Figura 26. Purificación de la proteína recombinante GPN en diferentes buffers:	65
Figura 27. Análisis con diferentes concentraciones de EDTA.	66
Figura 28. Análisis con diferentes concentraciones de DTT.	67
Figura 29. Análisis con diferentes concentraciones de aminoácidos.	68
Figura 30. Análisis con diferentes concentraciones de lisina.	69
Figura 31. Análisis con buffer empleando diferentes compuestos.	70
Figura 32. Concentración de proteína pura.	71
Figura 33. Análisis de dispersión de luz dinámica de la proteína GPN.....	73
Figura 34. Análisis preliminar de las imágenes de geles SDS-PAGE.....	75
Figura 35. Preprocesamiento de las imágenes de geles.	76
Figura 36. <i>Perfil de Imagen basado en segmentación de imágenes binarias.</i>	77
Figura 37. Análisis del nuevo perfil.	79
Figura 38. Análisis empleando la proteína BSA.....	80
Figura 39. Perfil obtenido del gel completo.....	81
Figura 40. Separación de la muestra control.	83
Figura 41. Perfil obtenido sobre la muestra control.....	83
Figura 42. Detección de bandas.	84
Figura 43. Correlación entre bandas detectadas automáticamente y la imagen de la muestra control.....	85
Figura 44. Gráfica para detectar peso molecular.	85
Figura 45. Cálculo de umbral.....	87
Figura 46. Muestra aleatoria para análisis.	88

Figura 47. Detección de GPN en base a su peso molecular.	88
Figura 48. Región de interés (ROI).	89
Figura 49. Perfil de zona ROI.	90
Figura 50. Áreas obtenidas de muestra GPN empleando diferentes metodologías.	93
Figura 51. Gráfica de datos normalizados.	95
Figura 52. Segmentación semántica aplicada para eliminar ruido de fondo.	95
Figura 53. Eliminación de ruido por cambio de modelo de color.	96
Figura 54. Cálculo de nuevo perfil sobre la segmentación semántica.	98
Figura 55. Cálculo del nuevo perfil sobre la región ROI obtenida de la segmentación semántica.	99
Figura 56. Aplicación del nuevo perfil sobre la imagen del gel completo.	99
Figura 57. Cálculo del nuevo perfil sobre las muestras de forma independiente.	101

B. Lista de tablas

Tabla 1. Tabla comparativa.	36
Tabla 2. Solubilidad obtenida por concentración salina.	64
Tabla 3. Datos del nuevo perfil.	80
Tabla 4. Proteína pura a diferentes concentraciones.....	81
Tabla 5. Margen de error en métodos de interpolación.	86
Tabla 6. Comparación de resultados aplicando diferentes metodologías.	91
Tabla 7. Matriz de confusión de muestras homogéneas.....	92
Tabla 8. Matriz de confusión de muestras heterogéneas.	92
Tabla 9. Exactitud obtenida de geles homogéneos y heterogéneos.....	92
Tabla 10. Normalización de resultados obtenidos.	94
Tabla 11. Valores obtenidos en cuatro imágenes de células que no han entrado en metástasis. Fuente: [68].....	102
Tabla 12. Valores obtenidos en cuatro imágenes de células que padecen metástasis. Fuente: [68].....	103
Tabla 13. Imágenes de melanoma de piel humana línea celular A375 y datos obtenidos. Fuente: [70]	104
Tabla 14. Progresión del carcinoma de células escamosas orales de hiperplasia epitelial a través de diferentes etapas de displasia. Fuente [71].	104
Tabla 15. Datos de los cortes elegidos al azar de los diferentes estados de displasia. Fuente: [71].....	105
Tabla 16. Tabla comparativa completa.....	114

Referencias

- [1] M. Faisal, T. Vasiljevic and O. N. Donkor, " A review on methodologies for extraction, identification and quantification of allergenic proteins in prawns," *Food Research International*, pp. 307-318, 2019.
- [2] A. Arbor and D. F. Keren, *Protein electrophoresis in clinical diagnosis*, NY: Oxford University Press, 2004.
- [3] M. Ferrari, L. Cremonesi, P. Carrera and P. Bonini, "Diagnosis of genetic disease by DNA technology," *Pure & Appl. Chem.*, vol. 63, no. 8, pp. 1089-1096, 1991.
- [4] A. Intarapanich, S. Kaewkamnerd, P. J. Shaw, K. Ukosakit, S. Tragoonrung and S. Tongsimma, "Automatic DNA diagnosis for 1D Gel Electrophoresis Images using Bio-image Processing Technique," *BMC Genomics*, pp. S12-S15, 2015.
- [5] L. Wai-Hoe, L. Wing-Seng, Z. Ismail and G. Lay-Ham, "SDS-PAGE-Based quantitative assay for screening of kidney stone disease," *Biological Procedures Online*, pp. 145-160, 2009.
- [6] B. Jania and K. Andraszek, "Application of native agarose gel electrophoresis of serum proteins in veterinary diagnostic," *J. Vet. Res.*, pp. 501-508, 2016.
- [7] M. M. Goetz, M. C. Torres-Madroño, S. Röthlisberger and E. Delgado-Trejos, "Preprocessing of 2-Dimensional gel electrophoresis images applied to proteomic analysis: A review," *Genomics proteomics bioinformatics*, pp. 63-72, 2018.

- [8] J. Choi, J. Hyun and S. Yang, "On-chip extraction of intracellular molecules in white blood cells from whole blood," *Scientific reports*, pp. 1-12, 2015.
- [9] S. Ahmed, A. Zahoor, M. Ibrahim, M. Yonnus, S. Nawaz, R. Naseer, Q. Akram, C. Deng and S. Chandra, "Enhanced efficacy of direct-acting antivirals in Hepatitis C patients by coadministration of black cumin and ascorbate as antioxidants adjuvants," *Oxidative medicine and cellular longevity*, pp. 1-10, 2020.
- [10] V. E. Pessoa, C. H. Alencar, M. T. Kamimura, F. Montenegro, S. G. De simone, R. F. Dtra and M. I. Florindo, "Occurrence of natural vertical transmission of Dengue-2 and Dengue-3 viruses in *Aedes aegypti* and *Aedes albopictus* in Fortaleza, Ceará, Brazil," *Plos one*, pp. 1-9, 2012.
- [11] T. S. Kim, K. M. Ho, K. R. Yim, W. S. Oh, S. B. Chon, S. Ryu, K. Yie and S. Lee, "Three reinfection cases of the pandemic influenza (H1N1 2009)," *Infect chemoter*, pp. 257-261, 2010.
- [12] E. González-González, G. Trujillo-de Santiago, I. M. Lara-Mayorga, O. M. Martínez-Chapa and M. M. Álvarez, "Portable and accurate diagnosis for COVID-19: Combined use of the miniPCR thermocycler and a well-plate reader for SARS-CoV-2 virus detection," *Plos one*, pp. 1-13, 2020.
- [13] R. Tao, Z. Ni, C. Liu, M. Zhu, X. Ji, X. Chen, J. Shen and S. Tu, "Expression, purification and identification of an immunogenic fragment in the ectodomain of prostate-specific membrane antigen," *Experimental and therapeutic medicine*, pp. 747-752, 2016.
- [14] H. Xu, T. Huang, Q. Yang, L. Xu, F. Lin, Y. Lang, H. Hu, Y. Peng, L. Tan, C. Qian and B. Huang, "Candidate tunir suppressor gene IRF6 is

involved in human breast cancer pathogenesis via modulating PI3K-regulatory subunit PIK3R2 expression," *Cancer management and research*, pp. 5557-5572, 2019.

- [15] T. Tomonaga, K. Matsushita, S. Yamaguchi, M. Oh-Ishi, Y. Kodera, T. Maeda, H. Shimada, T. Ochiai and F. Nomura, "Identification of altered protein expression and post-translational modifications in primary colorectal cancer by using agarose two-dimensional gel electrophoresis," *Clinical cancer research*, pp. 2007-2014, 2004.
- [16] J. G. Mohanty, H. D. Shukla, J. D. Williamson, L. J. Launer, S. Saxena and J. M. Rifkind, "Alterations in the red blood cell membrane proteome in alzheimer's subjects reflect disease-related changes and provide insight into altered cell morphology," *Proteome science*, pp. 8-10, 2010.
- [17] Instituto Nacional del Cáncer, "Estadísticas del cáncer," 20 04 2021. [Online]. Available: <https://www.cancer.gov/espanol/cancer/naturaleza/estadisticas>.
- [18] A. Romero-Utrilla, J. F. Osuna-Ramos, F. Candanedo-Gonzalez, M. G. Ramírez and J. E. Peñuelas, "Cáncer de mama: Entidad patológica de biología heterogénea," *Arch. Salud Sin.*, pp. 109-116, 2014.
- [19] International Agency for Research on Cancer (IARC) and World Health Organization (WHO), "Globocan 2020: Breast.," 2 07 2021. [Online]. Available: <https://gco.iarc.fr/today/data/factsheets/cancers/20-Breast-fact-sheet.pdf>.
- [20] American cancer society, "Understanding a breast cancer diagnosis," 21 11 2020. [Online]. Available: <https://www.cancer.org/content/dam/CRC/PDF/Public/8580.00.pdf>.

- [21] R. Barroso-Sousa and O. Metzger-Filho, "Differences between invasive lobular and invasive carcinoma of the breast: results and therapeutic implications," *Therapeutic advances in medical oncology*, pp. 261-266, 2016.
- [22] W. Truin, R. M. Roumen, S. Siesling, K. .: Van de Vijver, V. C. Tjan-Heijnen and A. C. Voogd, "Estrogen and progesterone receptor expression levels do not differ between lobular and ductal carcinoma in patients with hormone receptor-positive tumors," *Breast Cancer Res Treat*, pp. 133-138, 2017.
- [23] J. Yu, D. J. Dabbs, Y. Shuai, L. A. Niemeier and R. Bhargava, "Classical-Type invasive lobular carcinoma with Her2 overexpression," *Am J Clin Pathol*, pp. 88-97, 2011.
- [24] A. B. Hanker and C. J. Der, "The Roles of Ras family small GTPases in breast cancer," in *Handbook of Cell Signaling*, USA, Academic Press, 2010, pp. 2763-2772.
- [25] B. Humpries, Z. Wang and C. Yang, "Rho GTPases: Big players in breast cancer initiation, metastasis and therapeutic responses," *cells*, pp. 1-22, 2020.
- [26] Z. Li, R. Fang, J. Fang, S. He and T. Liu, "Functional implications of Rab27 GTPases in cancer," *ell Communication and signaling*, pp. 1-8, 2018.
- [27] B. Lara-Chacón and et al., "Gpn3 is essential for cell proliferation of breast cancer cells independent of their malignancy degree," *Technology in cancer research & treatment*, pp. 1-11, 2019.
- [28] A. B. Nowakowski, W. J. Wobig and D. H. Petering, "Native SDS-PAGE: High Resolution Electrophoretic Separation of Proteins With Retention

of Native Properties Including Bound Metal Ions," *Metallomics*, pp. 1068-1078, 2014.

- [29] M. H. Alnamoly, A. M. Alzohairy, I. Mahmoud and I. M. El-Henawy, "EGBIOIMAGE: A software tool for gel images analysis and hierarchical clustering," *IEEE Access*, pp. 10768-10781, 2019.
- [30] M. F. Abadi, "Processing of DNA and Protein Electrophoresis Gels by Image processing," *Science Journal*, pp. 3486-3494, 2015.
- [31] S. D. Bolboacă, "Medical Diagnostic Test: A review of test anatomy, phases, and statistical treatment of data," *Computational and Mathematical Methods in Medicine*, pp. 1-22, 2019.
- [32] K. R. Foster, R. Koprowski and J. Skufca, "Machine learning, medical diagnosis, and biomedical engineering research-commentary," *Bio-Medical engineering Online*, pp. 1-9, 2014.
- [33] N. Kaabouch, . R. R. Schultz and B. Milavetz, "An analysis system for DNA Gel Electrophoresis images based on automatic thresholding and enhancement," in *Electro/Information technology IEEE international*, 2007.
- [34] F. Cai, S. Liu, P. T. Dijke and F. J. Verbeek, "Image analysis and pattern extraction of proteins classes from one-dimensional gels electrophoresis," *International Journal of Bioscience, Biochemistry and Bioinformatics*, pp. 201-212, 2017.
- [35] X. Ye, C. Y. Suen, M. Cheriet and E. Wang, "A recent Development in Image Analysis of Electrophoresis Gels," in *Vision Interface '99, Trois-Rivières, Canada, Canada*, 1999.

- [36] L. Jian-Derr, H. Chung-Hsien, W. Neng-Wei and L. Chen-Song, "Automatic DNA sequencing for electrophoresis gels using image processing algorithms," *J. Biomedical Science and Engineering*, pp. 523-528, 2011.
- [37] R. S. Taher, N. Jamil, S. Nordin and U. M. Bahari, "A new false peak elimination method for poor DNA gel images analysis," in *International Conference on Intelligent Systems Design and Applications*, Okinawa, Japan, 2014.
- [38] R. Koprowski, Z. Wróbel, A. Korzynska, K. Chwialkowska and Kwasniewski, "Automatic analysis of 2D polyacrylamide gels in the diagnosis of DNA polymorphisms," *Biomedical Engineering*, pp. 1-15, 2013.
- [39] A. Galván, M. Tejada, A. Camargo, J. J. Hihuera and E. Fernández, "Transformación de Escherichia coli con un plásmido recombinante," 29 11 2022. [Online]. Available: <https://www.uco.es/dptos/bioquimica-biol-mol/pdfs/48%20TRANSFORMACION%20E%20COLI%20CON%20PLASMIDO%20RECOMBINANTE.pdf>.
- [40] J. Juárez-Lucero, M. R. G. Guevara-Villa, A. Sánchez-Sánchez, R. Díaz-Hernández and L. Altamirano-Robles, "Development of a Methodology to Adapt an Equilibrium Buffer/Wash Applied to the Purification of hGPN2 Protein Expressed in Escherichia coli Using an IMAC Immobilized Metal Affinity Chromatography System," *Separations*, pp. 1-16, 2022.
- [41] L. Coelho, A. Santos, H. Napoleao, M. Correia and P. Paiva, "Protein Purification by Affinity Chromatography," 28 11 2022. [Online].

Available: https://cdn.intechopen.com/pdfs/26596/InTech-Protein_purification_by_affinity_chromatography.pdf.

- [42] A. Efrat, F. Hoffmann, K. Kriegel, C. Schultz and C. Wenk, "Geometric algorithms for the analysis of 2D-Electrophoresis gels," *Journal of computational biology*, pp. 1-20, 2001.
- [43] I. Bajla, I. Holländer, S. Fluch, K. Burg and M. Kollár, "An alternative method for electrophoresis gel image analysis in the GelMaster software," *Computer Methods and Programs in Biomedicine*, pp. 209-231, 2005.
- [44] N. Labyed, N. Kaabouch, R. R. Schultz and B. B. Singh, "Automatic segmentation and band detection of protein images based on the standard deviation profile and its derivative," in *IEEE International Conference on Electro/Information Technology*, Chicago, 2007.
- [45] G. Ramaswamy, B. Wu and U. MacEvilly, "Knowledge management of 1D SDS PAGE Gel protein image information," *Journal of Digital Information Management*, pp. 223-232, 2010.
- [46] P. Tsakanikas, "Image processing methods and algorithms for accurate protein spot detection in 2-dimensional gel electrophoresis (2 DGE)," 22 11 2020. [Online]. Available: <https://www.semanticscholar.org/paper/Image-processing-methods-and-algorithms-for-protein-Tsakanikas/e89e0e1e4ecfdd8af4a54d6a31570ca259e5359f>.
- [47] D. Soto and P. Alvarado, "Automatic detection of bands in electrophoresis gel images by means of optimization of a target function," in *Conference on Technologies for Sustainable Development*, 2011.

- [48] S. Magdeldin, S. Enany, Y. Yoshida, B. Xu, Y. Zhang, Z. Zureena, I. Lokamani, E. Yaoita and T. Yamamoto, "Basic and recent advances of two dimensional- polyacrylamide gel electrophoresis," *Clinical Proteomics*, pp. 1-10, 2014.
- [49] J. M. Brauner, T. W. Groemer, A. Stroebel, S. Grosse-Holz, T. Oberstein, J. Wiltfeang, J. Kornhuber and J. M. Maler, "Spot quantification in two dimensional gel electrophoresis image analysis: comparison of different approaches and presentation of a novel compound fitting algorithm," *Bioinformatics*, pp. 1-12, 2014.
- [50] A. Abeykoon, M. Dhanapala, R. Yapa and S. Sooriyapathirana, "An automated system for analyzing agarose and polyacrylamide gel images," *Ceylon Journal of Science*, pp. 45-54, 2015.
- [51] J.-Z. Sim, P.-V. Nguyen, H.-K. Lee and S. K.-E. Gan, "GelApp: Mobile gel electrophoresis analyser," *Nature Methods Application Notes*, pp. 1-2, 2015.
- [52] M. Rezaei, M. Amiri, P. Mohajery and M. Rezaei, "A new algorithm for lane detection and tracking on pulsed field gel electrophoresis images," *Chemometrics and Intelligent Laboratory Systems*, pp. 1-18, 2016.
- [53] C. Fernández-Lozano, J. A. Seoane, M. Gestal, T. R. Gaunt, J. Dorado, A. Pazos and C. Campbell, "Texture analysis in gel electrophoresis images using an integrative kernel-based approach," *Scientific reports*, pp. 1-13, 2016.
- [54] M. K. Turan, A. Elen and E. Sehirli, "Analysis of DNA Gel Electrophoresis Images with Backpropagation Neural Network Based Canny Edge Detection Algorithm," *International Journal of Scientific and Technological Research*, pp. 55-63, 2016.

- [55] Q. Ou, J. Xiao, L. Yu, K. Wu and B. Xiong, "2D electrophoresis image brightness correction based on gradient interval histogram," *BMC bioinformatics*, pp. 1-11, 2020.
- [56] J. Juárez, A. Sánchez, R. Hernández, M. Guevara and L. Altamirano, "Identificación de bandas de proteína recombinante expresada en *Escherichia coli* usando una máscara binaria," in *4o. Congreso nacional de investigación interdisciplinaria*, ciudad de México, 2020.
- [57] N. Naganure and S. Kamath, "BEV Detection and Localisation using Semantic Segmentation in Autonomous Car Driving Systems," in *2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, 2021.
- [58] MathWorks, "3-D Brain Tumor Segmentation Using Deep Learning," 29 04 2021. [Online]. Available: <https://es.mathworks.com/help/vision/ug/segment-3d-brain-tumor-using-deep-learning.html>.
- [59] T. Wang, "Semantic Segmentation," 13 9 2021. [Online]. Available: http://www.cs.toronto.edu/~tingwuwang/semantic_segmentation.pdf.
- [60] J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation," *Computer Vision Foundation*, pp. 3431-3440, 2015.
- [61] D. Mwit, "The definitive guide to semantic segmentation for Deep Learning in Python," 30 4 2021. [Online]. Available: <https://cnvrg.io/semantic-segmentation/>.

- [62] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional Networks for biomedical image segmentation," 30 4 2021. [Online]. Available: <https://arxiv.org/pdf/1505.04597.pdf>.
- [63] A. Ghosh, A. Sufian, F. Sultana, A. Chakrabarti and D. De, "Fundamental Concepts of Convolutional Neural Network," 15 12 2022. [Online]. Available: https://www.researchgate.net/publication/337401161_Fundamental_Concepts_of_Convolutional_Neural_Network.
- [64] H. Gholamalinezhad and H. Khosravi, "Pooling Methods in Deep Neural Networks, a Review," 12 12 2022. [Online]. Available: https://www.researchgate.net/publication/344277235_Pooling_Methods_in_Deep_Neural_Networks_a_Review.
- [65] A. F. Agarap, "Deep Learning using Rectified Linear Units (ReLU)," 25 12 2022. [Online]. Available: https://www.researchgate.net/publication/323956667_Deep_Learning_using_Rectified_Linear_Units_ReLU.
- [66] R. González-González, J. A. Guerra-Moreno, G. R. Cristobal-Mondragón, V. Romero, S. G. Peña-Gómez, G. M. Montero-Moran and R. Sanchez-Olea, "Human Gpn1 purified from bacteria binds guanine nucleotides and hydrolyzes GTP as a protein dimer stabilized by its C-terminal tail," *Protein Expression and Purification*, pp. 85-96, 2017.
- [67] J. Juárez., M. R. G. Guevara-Villa, A. Sánchez-Sánchez, R. Díaz-Hernández and L. Altamirano-Robles, "Image Segmentation Applied to Line Separation and Determination of GPN2 Protein Overexpression for Its Detection in Polyacrylamide Gels," in *Progress in Artificial*

Intelligence and Pattern Recognition, Havana, Cuba, Springer, 2021, pp. 303-315.

- [68] J. M. Caron and J. M. Caron, " Methyl Sulfone Blocked Multiple Hypoxia- and Non-Hypoxia-Induced Metastatic Targets in Breast Cancer Cells and Melanoma Cells," *PLoS ONE*, pp. 1-22, 2015.
- [69] MacPride , "Breast Cancer Cells Vs Normal Cells Under Microscope," 8 september 2020. [Online]. Available: <https://microspedia.blogspot.com/2020/09/breast-cancer-cells-vs-normal-cells.html>.
- [70] S. Alarifi, D. Ali, S. Alkahtani, A. Verma, M. Ahamed, M. Ahmed and H. A. Alhadlaq, "Induction of oxidative stress, DNA damage, and apoptosis in a malignant human skin melanoma cell line after exposure to zinc oxide nanoparticles," *International Journal of Nanomedicine*, pp. 983-993, 2013.
- [71] J. B. Epstein and P. Güneri, "The adjunctive role of toluidine blue in detection of oral premalignant and malignant lesions," *Oral Health & Preventive Dentistry*, pp. 79-87, 2009.

Resumen en inglés

ABSTRACT

A methodology was developed that allowed the construction of a buffer to achieve the purification of the recombinant protein GPN expressed in the *Escherichia coli* bacterium, the method is based on the effect of the different additives that have been used and their influence on the links and/or solubility of both the contaminants and the protein itself.

The first factor that was analyzed was the growth temperature to obtain the protein of interest in a soluble form. The results indicated that to obtain the GPN within the soluble fraction it was necessary to perform growth at 10°C after having induced the expression of the recombinant protein with IPTG.

The second point to take account was to measure the solubility of the protein of interest within the proposed initial buffer based on Tris-HCl at a concentration of 100 mM to maintain the pH at 8.2. With the different concentrations of NaCl used to increase the solubility of the entire cell extract, it was found that at a concentration of 100 mM NaCl, the greatest amount of protein is achieved in its soluble form and that at higher values it begins to precipitate.

Taking the value of the saline concentration that provided a greater soluble fraction, variations were made in the same Tris-HCl buffer to identify its effect on the elimination of contaminants, from 100 mM of Tris-HCl no decrease in impurities was achieved by so far the buffer for the following experiments was worked at 100 mM Tris-HCl and 100 mM NaCl, which is the one that allowed to maintain the least amount of contaminants and the greatest amount of soluble GPN.

Subsequently, different concentrations of EDTA were used to reduce oxidation damage, inhibit metalloproteases, preventing the studied protein from denaturing, and measure its effect on the removal of impurities. These results

showed that for the GPN the use of 400 μ M EDTA helped to reduce the thickness of some of the protein bands considered contaminants while maintaining the same GPN concentration. The concentrations used were reduced since in large quantities nickel is released from the column, preventing IMAC chromatography.

As bands of the same thickness as the GPN continued to be present and could not be eliminated, it was concluded that they were possibly linked by disulfide bond on the GPN protein, so different concentrations of DTT were used to separate them. This reagent makes it possible to maintain the active state of the recombinant protein, eliminate disulfide bonds corresponding to contaminating proteins attached to the protein of interest, and measure its effect on their removal; the values used were small to avoid denaturing the GPN. The results indicated that the use of 100 μ M favors the elimination of contaminating proteins, being able to obtain the GPN protein in a pure form at a concentration of 2mg/ml.

To improve the solubility of the protein, the effect that different amino acids have on the purification process of the recombinant protein was sought, the use of Sodium Glutamate was intended to increase the ionic strength of the sample, stabilize the recombinant protein and decrease its production within inclusion bodies. Arginine was used for its similar properties to glutamate. However, the study carried out with arginine indicated that it does not significantly reduce contaminants and instead decreases the amount of GPN within the sample and eliminates it. A similar effect in the removal of contaminants was achieved using 100 mM Sodium Glutamate. To find a replacement for arginine that would maintain the same charge, but with less strength, the amino acid lysine was used since it is less polar than arginine. By using this amino acid at a concentration of 100 mM, it was possible to reduce the number of impurities and preserve a greater amount of the recombinant protein than when using arginine.

When the purification was carried out using only 100 mM Tris-HCl, 100 mM NaCl, 100 mM Lysine and 100 mM Sodium Glutamate at a pH of 8.2, it was possible to eliminate a considerable number of contaminants, maintaining two protein bands that continued to be preserved and that when performing the kinetics using DTT they could be eliminated. This suggests that these proteins were disulfide bonded on the GPN protein and therefore could be removed using DTT or 2-Mercaptoethanol which depends on the resin being used when performing the immobilized metal affinity chromatography.

With all the measurements carried out, it was determined that the ideal buffer to obtain the GPN protein in a pure form and at high concentrations should have the following reagents: 100 mM Tris-HCl, 5% glycerol, 0.2% Triton X-100, 100 mM NaCl, 400 μ M EDTA, 100 μ M DTT at pH 8.2. Glycerol was added to stabilize the buffer and prevent proteins from clumping. The Triton X-100 is used to prevent precipitation, solubilize insoluble proteins and release proteins present within inclusion bodies.

The results show that the use of this buffer created for the recombinant GPN protein expressed in *Escherichia coli* achieved, without the use of protease inhibitor tablets, the protein in a totally pure and homogeneous form, with low polydispersity corroborated by dynamic dispersion analyzes, achieving a concentration of 30 mg/ml so it can be used for crystallography or structural genomics studies.

The pure protein was concentrated to different volumes to emulate samples of different stages of Her2+ type breast cancer that is highly invasive classified as invasive ductal and lobular carcinoma (CDI and ILC). Electrophoresis of the different samples was performed, and SDS-PAGE gels were prepared to carry out the image analysis study.

382 gels were prepared with 10 samples in each of them, making a total of 3,820 samples, of which 2,259 were defective, discolored, or with little

expression of the GPN protein. 1,561 samples presented the expressed or overexpressed GPN protein, but with different shades of color, which is why they were called heterogeneous gels. From this sample of gels, 669 samples were taken that presented the same color conditions and were called homogeneous gels.

Controlled concentrations of the GPN protein were made with the values of 0, 2, 9.5, 14, 14.5, 18, 18.5, 26, 27, 30 mg/ml of solution. The concentrations were added to the bacterial extract to obtain samples that represented different expression of the recombinant protein and were representative of different stages of Her2+ type cancer.

The intensity profiles calculated on the images did not allow obtaining features that could identify either the number of samples per gel or the number of protein bands per sample, due to the presence of endogenous proteins within the same sample. Only in the case that there are no samples with contaminants (endogenous proteins), without background noise (same color without objects), and with well-defined proteins in the sample (as in the control sample) is it as with the profile of the sample. image could find the position of the protein bands. However, this is not the case when obtaining samples from patients, since there are always the proteins that the cell of the tissue that is being analyzed requires in order to carry out its metabolic functions.

Until now, no description of the use of computational tools was found to automatically analyze the number of samples per gel or any specific protein band, in the images of protein gels. In addition, none of the methods developed allow us to find the overexpression of any protein of biological interest. Therefore, it was decided to develop a methodology to carry out the analysis of images of protein gels starting with an image preprocessing. The colors of the grayscale images were inverted and binarized to achieve a better separation of the samples present in the gel using erosion and dilation techniques. Then, taking into account that in the separation of the samples or

protein bands in the gel image there must be an increase in the black color or, what is the same, a decrease in the white color, a technique was developed that allowed to find these variations in white color (related to the separation of the samples or of the protein bands) called "Image Profile Based on Segmentation of Binarized Images, IPBSBI".

The graph generated with the developed methodology made it possible to relate the multiple minima with the number of samples present in the gel. By choosing one of the samples and repeating the methodology developed in it, the graph that represented the new profile now allowed the multiple maxima to be related to the position of each of the protein bands present in the gel image.

The new image profile not only detected the number of samples present in the gel and the number of protein bands, but also allowed the size of each of the maxima to be related to the amount of protein expressed in the sample, that is, the larger the size, the higher the concentration. Its function was verified with a second protein, BSA, where the size of the local maxima in the graph corresponded to the amount of protein expressed.

The image profile based on segmentation of binary images, IPBSBI, allowed us to find the overexpressed amount of the GPN protein when the different concentrations of the protein were presented in the gel. However, the base of the local maxima did not start at the origin and a threshold had to be chosen that allowed the multiple maxima to be eliminated. This threshold caused the amount of samples present in the gel to be automatically detected.

Knowing the position of the number of samples that the gel has, the first sample that contains the control of the molecular weight of the proteins was chosen and the proposed methodology was used to detect the control proteins, know their molecular weight and develop a interpolation method that allowed knowing the size of the proteins present in the rest of the samples.

Having identified the molecular weight of the GPN protein, the region containing all the samples that include it at different overexpression was chosen, the image profile based on segmentation of binary images was applied and the size of the multiple maxima could be related to the amount of protein expressed per sample. This demonstrated that the methodology allowed detecting the overexpression of the protein of interest correctly as long as there is no background noise. If this condition is met, then the level of overexpression can be detected along with the order of expression.

The accuracy of the method was evaluated by means of the confusion matrix using homogeneous gels, that is, they presented the same color, without distortions and a value of 0.985052 was obtained. The analysis was repeated using all the samples including the heterogeneous gels and the accuracy value dropped to 0.91736.

The proposed methodology was compared with other methods that included calculating the areas of the protein samples manually, using K-means segmentation and Otsu segmentation. However, the binary image segmentation-based image profiling provided better results in finding the order of GPN protein overexpression than the remaining methods did not. In addition, it was necessary for each of the methods used to have each of the images containing the GPN protein with different concentrations, since there is no methodology that allows analyzing these samples within an SDS-PAGE protein gel.

To improve the proposed methodology, semantic segmentation was applied to eliminate background noise and detect only the proteins present in the gel, it was complemented by transforming the images obtained from the RGB color space to HSV and a color filter was made that allowed to eliminate by Complete the unwanted proteins in each of the samples. In these images, the image profile based on segmentation of binary images was applied again, obtaining a new noise-free graph where the size of the local maxima was related to the

amount of protein expressed within each sample and the order of overexpression was detected. of the protein.

After corroborating that the image profile based on segmentation of binary images made it possible to identify the thickness or size of the bands, the possibility of using it to identify the presence of large cell numbers related to different types of cancers was raised, since this condition is It is also related to the increase in the number of affected cells, it was verified if the developed methodology has the capacity to detect said cell increase. This approach arises because by increasing the number of cells and having the binary image and with inverted colors, the spaces between them must decrease because the cell number is increasing, this would be related to the increase in the white color within the picture.

Images of cell samples were taken where the number of cells obtained at the time of applying a solution that inhibits the growth of breast cancer cells is visually analyzed. Binary image segmentation-based image profiling was able to detect the decrease in cell number by relating the image to the size of the peak on the plot. The values found were in agreement with the state in which the cells were found, that is, in samples where there are no affected cells, the values of the size of the maximums were lower than in the cases where there are cancer cells.

The same results were found when analyzing different types of cell samples. For this reason, the image profile based on segmentation of binary images allowed not only to detect protein overexpression in SDS-PAGE gels, but also had the ability to detect the increase in the number of cells within the image.

The developed programs Scanalytics, GelcomparII, Gel-Pro Analyzer, TotalLab, PDQuest, Proteomweaver, Dcyder 2D, imageMaster, Melanie, BioNumerics, Redfin, Gel IQ, Z3 and Delta2D Flicker are directed to the analysis of DNA gel images, in the identification of bands in protein gels,

different filters are applied to eliminate as much noise as possible. Furthermore, they are all semi-automatic, they require an operator to interact with the program when choosing a column or band of interest. The user also adjusts the intensity changes and decides the best threshold that contains less background noise. In addition, some of the programs carried out are very complicated for analysts who do not have knowledge about intensities and thresholding, consequently, they do not get a good analysis of the gel.

The programs developed, the most complete are those of Alnamoly, GelAnalyzer and JJuárez (obtained with this studies). Of these, only Alnamoly and the proposal made in this work can detect both the columns and the bands of the gel. However, the proposal of this doctoral research does not require the user to place the name of the samples surpassing Alnamoly, in addition it does not require grouping the bands to be able to detect them, it can find the protein overexpression which none of the above does. and so far, it has only been used for protein gels and not for DNA.

Thus, finally, image profiling based on binary image segmentation can detect the level of overexpression of protein bands within an SDS-PAGE gel, it can identify the number of samples and the number of protein bands. . In addition, it could identify if a sample obtained from a tissue that contains a cancer-related condition is getting worse or better, so it can be used as a reliable diagnostic method, since if the samples are analyzed manually by laboratories, they can be confused. and fail to detect if the disease is advancing or if the patient is showing some improvement.

Therefore, the developed technique can identify which sample contains a higher protein concentration, so it can be used as a new diagnostic method to identify if a patient is in an aggressive stage of Her2+ type breast cancer. It can even be used to identify if the use of any therapy or medication is working to improve the health of the patient, since if the GPN protein is decreasing in its

concentration, it is synonymous with the fact that the patient has improvements over the disease.

The methodology was applied to compare images that show normal breast cells and cancer cells, being able to detect how metastasis is increasing. It was also tested with images of human skin melanoma in A375 cell lines where the increase in cell damage could be verified.

Finally, binary image segmentation-based image profiling was able to identify carcinoma progression at different levels of epithelial hyperplasia by analyzing the entire image and cutting out small 25x35 pixel regions.

All these results indicate that the proposed methodology can be used to identify the overexpression of proteins of biological interest and thus detect in samples from cancer patients if the disease is advancing or even if some treatment is working as long as the type of cancer is present. It is related to the overexpression of a specific protein. Binary image segmentation-based image profiling can also be used to identify in images of cellular tissues whether the number of cells is increasing which is a sign that the cancer is increasing.

The results obtained indicate that the developed methodologies can be used as a diagnostic method to find the overexpression of a protein of biological interest related to a cancer condition and how it can also be used to detect the different stages of different types in cell images. of cancer, the proposed objectives created from the research question have been exceeded.

Therefore, after the analysis carried out, the objectives of the thesis were achieved by responding positively to the research hypothesis, as well as to the research question, obtaining a system that can be used as a diagnostic method to identify the overexpression of related proteins. with breast cancer. This technique can be used not only to detect the GPN protein or some other GTPase, but it can also be used to detect the overexpression of other types of proteins related to different types of cancers, such as prostate cancer.

Additionally, this methodology turns out to be cheaper to detect proteins considered biomarkers of diseases than when looking for them using molecular techniques.

This research can produce a methodology that is contemplated to create a patent, and a various papers so national and international.