CARACTERIZACIÓN MOLECULAR Y ANOTACIÓN FUNCIONAL DE UNA PROTEÍNA HIPOTÉTICA (GENBANK: CAI46211.1) DEL HOMO SAPIENS.

Tensile Strength of Materials
Pontificia Universidad Javeriana, Bogotá, Colombia
Laura Echeverry, Paula Ugueto, William Gómez

***Index Terms*—Proteína hipotética, Análisis bioinformático, Tejido cerebral, C-quinasa, Homo Sapiens.**

I. INTRODUCCIÓN

Esta proteína hipotética fue secuenciada por el Centro de Investigación Alemán de Cáncer junto con la Universidad Ludwig Maximilians en el marco del Proyecto del Genoma Alemán [1]. El principal autor de esta secuenciación es el Profesor Dr. Stefan Wiemann, que ha enfocado sus trabajos en el cáncer de seno. El cáncer, así como muchas otras enfermedades humanas nace de aberraciones genéticas que pueden ser heredadas o pueden ocurrir espontáneamente en células somáticas. Estos defectos causan actividades anormales en los productos de genes y provocan funcionamientos erróneos en interacciones moleculares y celulares, terminando en tumores y progreso del cáncer. El objetivo central de esta unidad de investigación es entender la complejidad de los mecanismos moleculares en la regulación de redes de señalamiento y cómo esto impacta en el desarrollo del cáncer, metástasis y la resistencia a las drogas. Para esto se hacen proyectos a gran escala, usando tecnologías genómicas y proteómicas para analizar diferentes genes y proteínas candidatos a influenciar en el tema. Por medio de estos análisis se construye conocimiento que se explota posteriormente para la identificación de nuevos marcadores para diagnóstico y pronóstico, así como para desarrollar estrategias de intervención terapéutica. Es así como el estudio de estas proteínas hipotéticas da luces sobre aquellas que están involucradas en los procesos de metástasis y cáncer de mama, estudiando sus actividades intrínsecas y sus redes con otras proteínas o genes. Los estudios han demostrado que las redes proteicas asociadas a tumores cancerígenos son complejas y que involucran diferentes tipos de células, incluyendo aquellas ubicadas en el microambiente del tumor y aquellas dentro de. Se puede, por lo tanto, estudiar sobre el impacto que tienen perturbaciones individuales a diferentes niveles (ADN, ARN, proteínas, metabolitos, fenotipos) en una gran variedad de vías celulares. Con esto se busca tener un mejor entendimiento de la conectividad entre sistemas de interacción en múltiples capas. Actualmente, de la proteína hipotética CAI46211 se sabe que tiene una longitud de 934 aminoácidos que se encuentra en el ser humano (Homo Sapiens). El tejido en el que fue hallada es el cerebral y en etapa de desarrollo fetal. Está altamente relacionada, como veremos a lo largo de este artículo, con una proteína quinasa de unión tipo C. Por lo tanto, hace parte de una familia de proteínas quinasas enzimas (PKC) involucradas en controlar la función de otras proteínas por medio de la fosforilación de los grupos hidroxilos en los residuos de aminoácidos serina y treonina en estas proteínas. Por lo tanto, las enzimas PKC juegan un papel importante en numerosas cascadas de señales de transducción. Con el fin de caracterizar una proteína hipotética haciendo uso de herramientas bioinformáticas se buscó en la base de datos de proteínas de NCBI una proteína hipotética que fuera de nuestro interés para hacer un análisis de caracterización molecular y la anotaciones funcionales de esta proteína, con el fin de identificar su función, ubicación, familia, estructura, dominios, motivos principales y sus

II. METODOLOGÍA

Sequence retrival and similarity identification La secuencia de la proteína hipotética fue tomada de la base de datos de NCBI. Se descargó en formato FASTA para poder buscarla en diferentes plataformas que nos acerquen a su caracterización. El primer paso para inferir la función de esta proteína fue buscar similitudes con las herramientas del NCBI, se corrió un BLAST para identificar similitudes.

Multiple sequence alignment and phylogeny analysis Se implementó MUSCLE, herramienta del EBI para un análisis comparativo y filogenético entre nuestra proteína hipotética y cinco proteínas homólogas. Adicionalmente con Phylogeny.fr se complementó el alineamiento múltiple.

Physiochemical properties analysis

Se determinaron las propiedades físico-químicas de la proteína por medio de la plataforma ProtParam, una herramienta que suministra ExPASy. Estas incluyen el peso molecular, el pI teórico, los principales aminoácidos que la componen, su composición atómica, la media vida estimada, su índice de inestabilidad y su índice alifático, el número total de residuos cargados negativa y positivamente y las predicciones de GRAVY (grand average hydropathicity). **Subcellular localization analysis** Se utilizó DeepLoc para predecir su ubicación, esta parece ser el núcleo. Los resultados se validaron con WoLFPSORT, que arrojó como segunda opción de ubicación citosol-núcleo.

Conserved domain, motif, fold, coil, family, and superfamily identification Se corrió la secuencia de la proteína para encontrar en el banco del NCBI algún dominio conservado que pueda indicar las funciones principales de la proteína, así como la familia y superfamilia a la que puede pertenecer. Adicionalmente, se buscó en MOTIF de GenomeNet los motivos identificables en la proteína y utilizando InterPro se identificaron sus términos de Gene Onthogoly, confirmaron presencia de dominos conservados y familias homólogas.

Secondary structure prediction La plataforma de PSIPRED nos permitió visualizar la estructura que conforma cada uno de los aminoácidos de la proteína. Del mismo modo, SOPMA arroja información sobre las proporciones de estructuras que se tienen en base a la cantidad de aminoácidos que las conforman.

Three-dimensional structure prediction Al haber hallado en las herramientas previas homologías entre esta proteína hipotética y otras, se utilizó el programa Swiss Model para predecir su estructura tridimensional.

ESCRIBIR LO QUE SE HIZO EN ALPHAFOLD

Modal queality assessment

Active site detection

III. Resultados y Discusión

El flujograma del estudio se muestra en la Figura 1

A. *Sequence and similarity information*

TABLA 1

Description	Organism	Query cover	E value	% identity	Access (NCBI)
protein kinase C-binding protein 1 isoform CRA_e	Homo sapiens	97%	0.0	99.56%	EAW 75707.1
protein kinase C-binding protein 1 isoform CRA_k	Homo sapiens	97%	0.0	99.56%	EAW 75714.1
protein kinase C-binding protein 1 isoform X7	Pan troglodytes	92%	0.0	96.18%	XP_016 793551.2
protein kinase C-binding protein 1 isoform X3	Pongo abelii	92%	0.0	95.96%	XP_024 094347.1
protein kinase C-binding protein 1 isoform X3	Macaca thibetana	92%	0.0	95.17%	XP_050 601089.1

B. *Multiple sequence alignment and Phylogenetic analysis*

TABLA 2

Fig. 1. Flujograma del trabajo

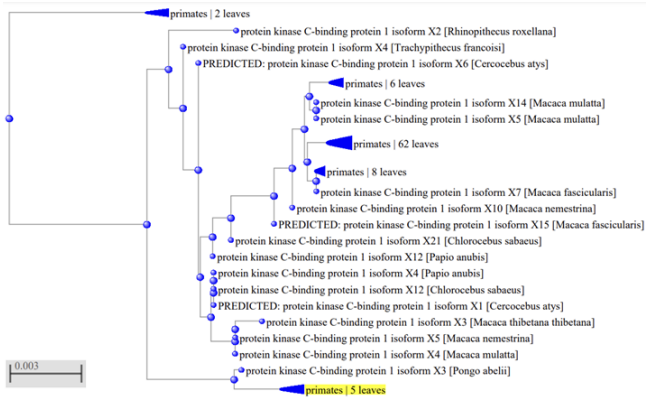

Fig. 2. Árbol con resultados de proteínas similares a la proteína hipotética en cuestión arrojados por Blastp

C. *Physicochemical features*

TABLA 3

Identity matrix of the hypothetical protein $(tr\|Q5JV90\|Q5JV90_HUMAN)$and its 5 most homologous proteins						
XP_024 094347.1	100.00	99.03	93.42	91.64	93.36	93.79
XP_050 601089.1	99.03	100.00	92.69	91.00	92.61	93.04
XP_016 793551.2	93.42	92.69	100.00	91.85	93.58	94.00
EAW 75714.1	91.64	91.00	91.85	100.00	97.53	97.86
tr—Q5JV90 —Q5JV90 _HUMAN	93.36	92.61	93.58	97.53	100.00	99.57
EAW 75707.1	93.79	93.04	94.00	97.86	99.57	100.00

Fig. 3. Figura 3.

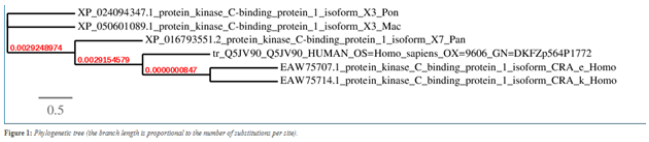

Fig. 4. Figura 4

D. Functional annotation of the hypothetical protein

E. Subcellular localization nature

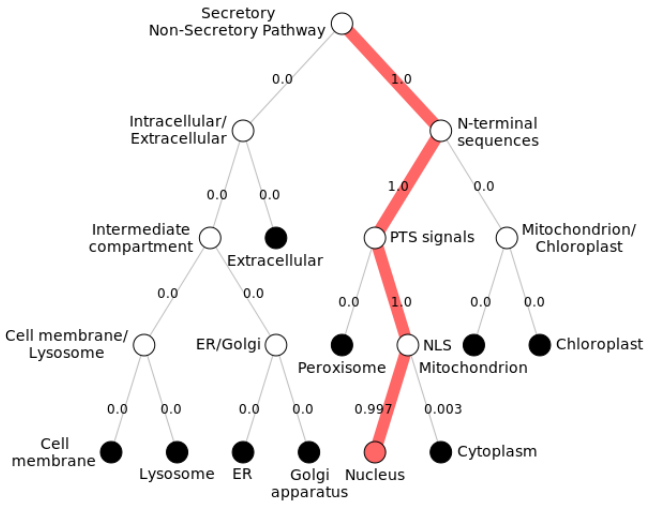

Fig. 5. Ubicación celular de la proteína hipotética

Amino acid	Amount	Percentage
Ser (S)	101	10.8%
Lys (K)	90	9.6%
Pro (P)	77	8.2%
Glu (E)	66	7.1%
Thr (T)	61	6.5%

id	site	distance	identity
UN33_CAEEL	cyto	70.7743	11.3127%
TAF5_DROME	nucl	72.3422	12.955%
PQE1_CAEEL	nucl	72.5807	12.8641%
LAF4_HUMAN	nucl	74.219	13.4365%
TAF2_DROME	nucl	75.2537	13.0221%
RSG3_HUMAN	plas	75.9928	13.738%
CC27_HUMAN	nucl	76.7311	12.8342%
TC2N_MOUSE	nucl	80.1314	13.2762%
NSD1_HUMAN	nucl	83.01	10.4228%
EG44_CAEEL	nucl	83.3044	11.9914%
SKI_HUMAN	nucl	85.2111	14.8663%
ELL_HUMAN	nucl	85.5608	13.0621%
NFX1_HUMAN	nucl	85.9171	12.7717%
GEM5_HUMAN	cyto_nucl	86.0225	12.7984%

Fig. 6. Localizacion celular de WoLFPSORT

Details	Protein ID	Score	Expected Accuracy	Localization Class	Gene Ontology Terms	Annotation Type
Details	CAI46211.1	24	84%	nucleus	histone acetyltransferase complex GO:0000123(IDA);	PSI-BLAST

Fig. 7. Ubicacion celular de LocTree3

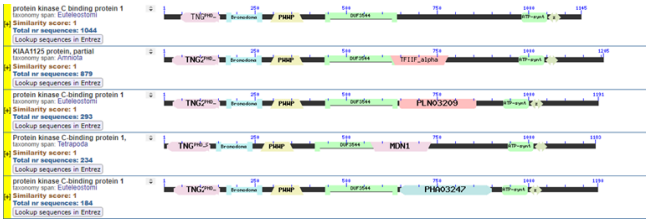

Fig. 8. Dominios conservados

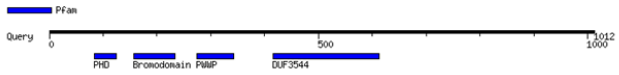

Fig. 9. Motivos con InterPro y Pfam

F. Conserved domains, motifs, family and superfamily identification

G. Secondary structure analysis

TABLA 4

H. Three-dimensional structure analysis

En esta homología se encuentran las similitudes en cuanto a las hélices Alpha. Sin embargo, se observa que los extremos y algunas zonas no plegadas en forma de hélice difieren en cuanto al modelo elegido. Por lo mismo, el programa arroja

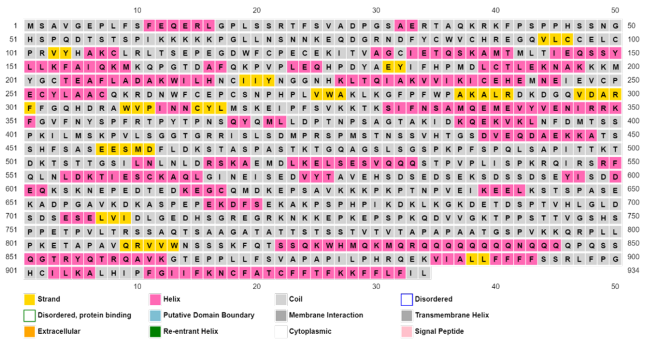

Fig. 10. Configuración de los diferentes aminoácidos que conforman la proteína

Alpha Helix	222	23.66%
Extended strand	129	13.81%
Random coil	555	59.42%
Beta turn	29	3.1%

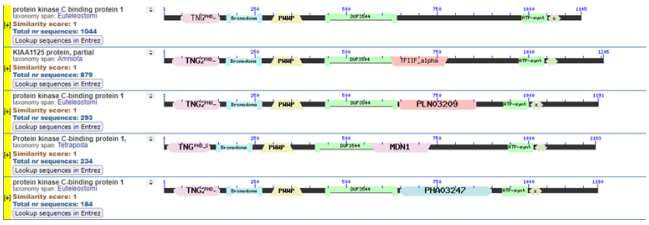

Fig. 11. Dominios conservados

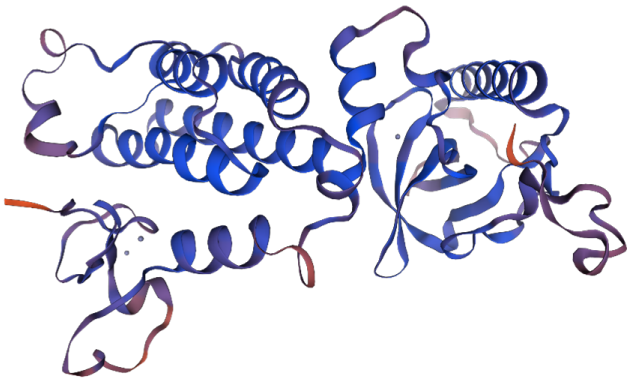

Fig. 12. Estructura tridimensional predicha en base a homologías

otras dos homologías: Transcription intermediary factor 1-alpha (código de acceso 3o35.1.A) y bromodomain PHD finger transcription (código de acceso 2f6j.1.A). Con una identidad del 30.72

Sin embargo, se observa en la Figura que una gran parte de la proteína no está modelada, puesto que no se encontraron homólogos que correspondieran a estas zonas. La línea verde es la secuencia de aminoácidos de la proteína hipotética, y toda la zona que no tiene azul por debajo es la que no fue modelada.

El resultado que arroja AlphaFold para la estructura tridimensional de la proteína tienen niveles de certeza que corresponden a los resultados de Swiss Model.

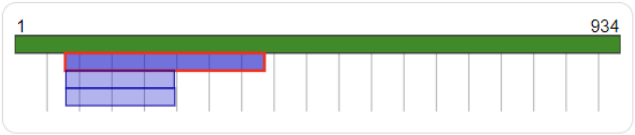

Fig. 13. Coverage of the three-dimensional structure

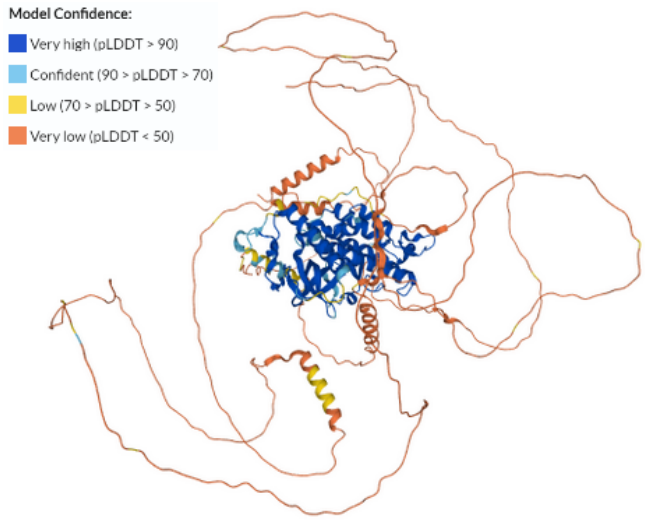

Fig. 14. Confidence of the three-dimensional AlphaFold2 structure

Como se observa en la Figura 14, las zonas de confianza alta son las mismas que aparecen en el modelo anterior, mientras que las de confianza muy baja son las que Swiss Model no modeló por falta de homologías.

Para corroborar la estructura 3D de la proteína hipotéitca tambien se utilizó phyre2,

El resultado de esta herramienta es el que se muestra en la imagen a continuación, y se reporta 100% confidencia en una Cobertura del 33% de la proteína. Además, se obtiene el tamaño en Armstrong de las dimensiones de la proteína.

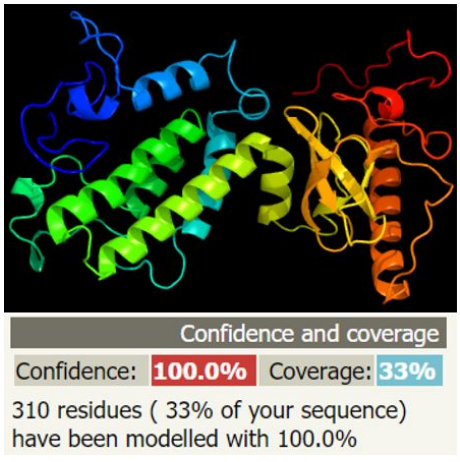

Fig. 15. Three-dimensional Phyre2 structure

En PROCHECK y Verify3D se corroboraron los resultados de los modelos obtenidos en las herramientas previamente

mencionadas, sin embargo, el resultado no fue positivo. Como se observa en la Figura, apenas 31.48% de los residuos tuvieron un score 3D-1D superior o igual a 0.2. Como este porcentaje es inferior al 80%, no se considera que sea un buen modelo.

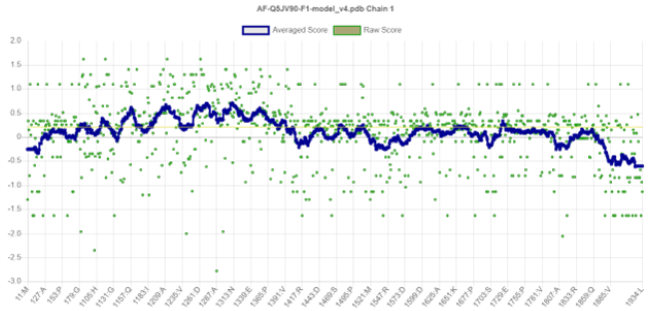

Fig. 16. Three-dimensional Phyre2 structure

Por otro lado, PROCHECK arrojó los siguientes resultados: el valor del Ramachandran plot es del 64%, por lo tanto, no alcanza el porcentaje de un modelo válido. Como se observa en la Figura , 17.6

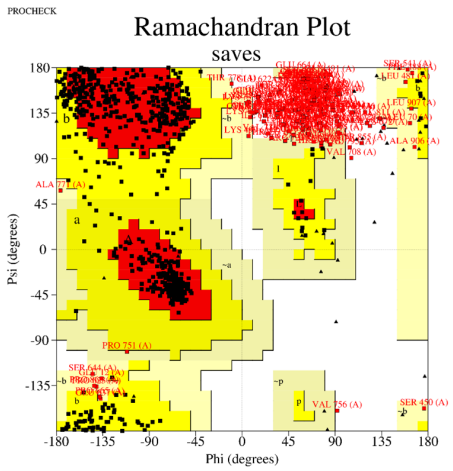

Fig. 17. Three-dimensional Phyre2 structure

Estos resultados desfavorables de los modelos tridimensionales de la proteína se deben a que las homologías existentes solo cubren un porcentaje de la proteína hipotética. El resto de la misma no tiene otras proteínas para compararla, por lo que se tiene una cobertura de apenas el 33% de la misma . La predicción que hace AlphaFold del resto de la proteína es igual de muy baja confianza.

I. Protein-protein interaction analysis

Cuando se corrió STRING, arrojó una proteína con un 94,5% de similitud: una proteína quinasa C-binding protein (código de acceso ZMYND8), que puede actuar como un correpresor transcripcional para KDM5D. Requerida para la regulación de diversos genes asociados con metástasis. Además, involucrada en la supresión de invasión celular en el cáncer de próstata. Por otro lado, contiene el bromodominio.

Las interacciones halladas de esta proteína son las que se muestran en la Imagen 18.

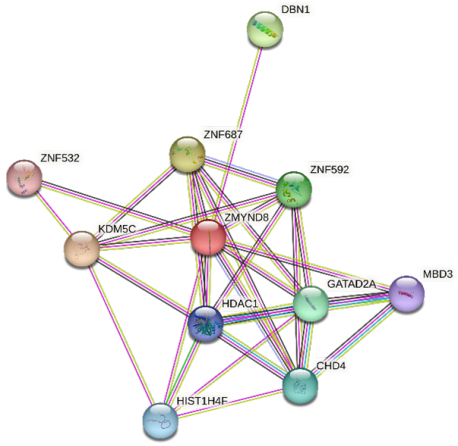

Fig. 18. Interacciones de la proteína ZMYND8, homóloga a la proteína hipotética en cuestión. Verde: gen vecino, Rojo: Fusión de genes, Azul: gen con co-ocurrencia, Azul claro: información de bases de datos curadas, Fucsia: información determinada experimentalmente.

J. Active site of the hypothetical protein

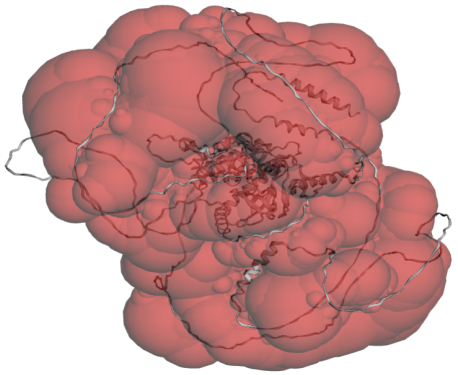

Fig. 19. Interacciones de la proteína ZMYND8, homóloga a la proteína hipotética en cuestión. Verde: gen vecino, Rojo: Fusión de genes, Azul: gen con co-ocurrencia, Azul claro: información de bases de datos curadas, Fucsia: información determinada experimentalmente.

IV. Conclusion

La identificación de funciones proteicas, sobre todo en el genoma humano, es fundamental para el mejor entendimiento de procesos biológicos. Lo que puede tener impactos positivos en aplicaciones biotecnológicas y médicas. Este estudio, enfocado en identificar posibles funciones, conformaciones e interacciones, implementó múltiples herramientas bioinformáticas con las cuales se caracterizó, encontrando varias proteínas homólogas, la proteína hipotética CAI46211.1. Se encontraron funciones de unión para quinasas tipo C, actividad transductora, así como 4 motivos claramente identificados: DUF3544, Bromodominio, PHD, y PWWP. Por lo que se resalta la importancia de esta proteína hipotética en sus posibles repercusiones sobre los estudios de cáncer y se espera que se siga el análisis de esta proteína que podría influenciar los estudios de terapia de cáncer con inhibidores de bromodominios.

Protein	Name	Possible function
KDM5C	Desmetilasa 5C lysine specific	Histona desmetilasa que desmetila específicamente 'Lys-4'
ZNF687	Zinc finger protein 687	Puede estar involucrada en regulación transcripcional
DBN1	Drebrin 1	Drebrinos pueden jugar algún papel en migración celular, extensión de procesos neuronales y plasticidad de las dendritas
ZNF592	Zinc finger protein 592	Puede estar involucrada en regulación transcripcional
GATAD2A	Transcriptional repressor p66-alpha	Optimiza la represión mediada por MBD2
CHD4	Chromodomain -helicase-DNA -binding protein 4	Componente del complejo histona desacetilasa NuRD que participa en el remodelamiento de la cromatina desacetilando histonas
HIST1H4F	Histone cluster 1 H4 family member f; Core component of nucleosome	Las histonas juegan un rol central en la regulación transcripcional, reparación del ADN, replicación del ADN y la estabilidad del cromosoma. Los nucleosomas envuelven y compactan el ADN en cromatina.
HDAC1	Histone deacetylase 1/2	Responsible de desacetilar residuos de lisina en la parte N-terminal de las histonas centrales (H2A, H2B, H3 y H4). La desacetilación de histonas da un tag para la represión epigenética y juega un rol importante en la regulación transcripcional, la progresión del ciclo celular y el desarrollo de eventos.
MBD3	Methyl-CpG -binding domain protein 3	Actúa como un represor transcripcional y juega un rol en el silenciamiento de genes.
ZNF532	Zinc finger protein 532	Puede estar involucrada en regulación transcripcional

REFERENCES

[1] N. Özkaya, D. Leger, D. Goldsheyder y M. Nordin, Fundamentals of Biomechanics Equilibrium, Motion, and Deformation, Switzerland: Springer, 2018.

[2] areatecnologia., area tecnologia, [En línea]. Available: https://www.areatecnologia.com/materiales/ensayo-de-traccion.html. [Último acceso: 02 11 2022].

[3] E. Lawrence, INSTRON, [En línea]. Available: https://www.instron.com/en/testing-solutions/iso-standards/iso-527-2. [Último acceso: 02 11 2022].

[4] M. J. C. Loaiza, P. G. Díaz y e. al, Influencia de la posición de impresión y la densidad de relleno, Revista Ingenierias Universidad de Medellín, vol. 2, p. 15, 2020.

[5] UMBC, UMBC, MicroMaterials Characterization Lab, [En línea]. Available: https://mmc-lab.umbc.edu/resources/instron/. [Último acceso: 02 11 2022].

[6] Travieso-Rodriguez JA, Jerez-Mesa R, Llumà J, Traver-Ramos O, Gomez-Gras G, Roa Rovira JJ. Mechanical Properties of 3D-Printing Polylactic Acid Parts subjected to Bending Stress and Fatigue Testing. Materials (Basel). 2019 Nov 22;12(23):3859. doi: 10.3390/ma12233859. PMID: 31766653; PMCID: PMC6926899.