

Predicción de la estructura secundaria y terciaria de GNL3L mediante herramientas computacionales

Prediction of the secondary and tertiary structure of GNL3L via computational tools.

Juan Felipe Cardona Arango, Sarah Röthlisberger

Ingeniería Biomédica, Instituto Tecnológico Metropolitano, Medellín, Colombia

Correo-e: juancardona120696@itm.edu.co, sarahrothlisberger@itm.edu.co

Resumen— GNL3L es una proteína poco caracterizada perteneciente al grupo de las Nucleosteminas. Sus funciones han sido asociadas principalmente a la génesis de los ribosomas, aunque recientes investigaciones la han asociado con la regulación del ADN telomérico, actuando como proteína accesoria facilitando la formación del complejo shelterina que protege el telomero. Actualmente no existen estudios experimentales que hayan definido la estructura cristalina de GNL3L, lo cual es importante para entender cómo actúa en el proceso de regulación de la longitud de los telómeros. En este trabajo, mediante herramientas computacionales (GOR, SOPMA, PSIPRED) se predice que GNL3L está compuesta en promedio de 52% de hélices alfa y 10% extended strands, lo cual permite inferir que la proteína es principalmente helicoidal. Se obtuvieron 5 modelos de estructura terciaria usando el programa MODELLER, de los cuales se optimizó el mejor con 3Drefine, se validó y graficó con ProSA-web.

Palabras clave— Análisis computacional de secuencias, bioinformática, estructura proteica, homología.

Abstract— GNL3L is a partially characterized protein, belonging to the Nucleostemin family. Its function is mainly in ribosome biogenesis, but recent investigations have shown that it is also involved in the regulation of telomeric DNA, acting as an accessory protein to facilitate the formation of the shelterin complex, which protects the telomere. To date there is no published experimental data on the crystallographic protein structure of GNL3L, which is important for understanding its role in telomere maintenance. In this work computational tools (GOR, SOPMA, PSIPRED) were used to predict that GNL3L is composed on average of 52% alpha helices and 10% extended strands, concluding that it is a mainly globular protein. Five models of GNL3L potential tertiary structure were obtained using the program MODELLER; of these the best was optimized with 3Drefine, validated and visualized with ProSA-web.

Key Word —Bioinformatics, computational sequence analysis, homology, protein structure.

I. INTRODUCCIÓN

Los telómeros son estructuras de ADN altamente especializadas localizadas al final de los cromosomas. Estas estructuras son cruciales para estabilizar los cromosomas y protegerlos de la fusión de sus extremos y la degradación. En el ser humano los telómeros están compuestos por un tándem repetitivo (TTAGGG) que está asociado con unas proteínas protectoras (TRF1, TRF2, TIN2, POT1, PIP1 y POT1) que comprenden el complejo shelterina [1][2]. Los telómeros se acortan progresivamente durante cada división celular hasta llegar a una longitud en la cual son incapaces de ser protegidos por el complejo shelterina; en este punto la célula responde desencadenando los mecanismos de senescencia y apoptosis [3]. Durante el proceso patológico del cáncer la telomerasa, un complejo multiproteico, se expresa añadiendo repeticiones al final de los telómeros evitando así el acortamiento de estos y por tanto induciendo una proliferación indefinida de las células. Si bien la telomerasa es expresada en algunas células somáticas en condiciones normales, en las células cancerígenas su nivel de expresión y actividad es más alta [4]. Se ha evidenciado que la proteína TRF1, del complejo shelterina, modula la longitud de los telómeros, con el fin de evitar una proliferación celular indefinida, a través de su interacción con las proteínas TIN2, TPP1, POT1 y PINX1 que actúan para inhibir la actividad de la telomerasa [5]. Hace no mucho se descubrió una nueva proteína llamada Guanine nucleotide binding protein-like 3-like (GNL3L) que se une a la proteína TRF1 del complejo shelterina, promoviendo así la homodimerización de TRF1 y su unión al telómero, y además promueve la estabilidad de esta proteína, inhibiendo su ubiquitinación y degradación [6]. Se propone que de esta forma se regula el acceso de la telomerasa al telómero para evitar el alargamiento anormal de los telómeros. A pesar de que GNL3L es una proteína importante en la regulación de la longitud de los telómeros, no existen estudios experimentales que hayan definido la estructura cristalina de GNL3L, ni su mecanismo de interacción con TRF1, lo cual es importante para entender el proceso de regulación de la longitud de los telómeros. Según la revisión realizada de la literatura, solo se encontraron 7 artículos

relevantes de GNL3L en bases de datos como PubMed y Science Direct [7][8][6][9][10][11][1], los cuales enfatizan la función de GNL3L en la regulación de la longitud de los telómeros. El presente trabajo tiene como objetivo presentar la predicción de la estructura secundaria y terciaria de GNL3L mediante herramientas computacionales, para así aportar al conocimiento que se tiene de su función.

II. METODOLOGIA

La metodología usada para la predicción de la estructura cristalina de GNL3L se resume en la Figura 1 y se describe a continuación.

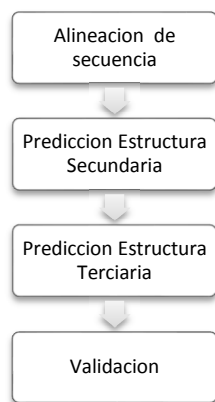


Figura 1: Diagrama de flujo de la metodología implementada

La secuencia de aminoácidos de GNL3L fue obtenida del National Centre of Biotechnology Information (NCBI), en formato FASTA, con código de acceso NP_001171748.1.

A. Alineación de secuencias

Con el fin de encontrar proteínas homólogas a GNL3L, con estructura cristalina ya determinada, la secuencia de aminoácidos fue sometida a herramientas BLAST (BLASTp, PSI-BLAST+ y Protein BLAST+), usando los parámetros predeterminados y haciendo búsqueda en la base de datos Protein Data Bank (pdb).

B. Predicción de la estructura secundaria

Para predecir la estructura secundaria de GNL3L, la secuencia de aminoácidos fue analizada en programas de predicción de estructura secundaria como PHYRE2, PSIPRED, GOR y CFSSP, disponibles en ExPASy (http://www.expasy.org/proteomics/protein_structure). Los resultados obtenidos de estos programas fueron tabulados y comparados con el fin de encontrar una estructura consenso.

C. Predicción de la estructura terciaria

Para la predicción de la estructura terciaria se empleó el programa MODELLER (<http://toolkit.tuebingen.mpg.de/modeller>) [13], el cual generó diversos modelos de la estructura terciaria de la proteína a partir de la alineación de varias proteínas similares a GNL3L.

Los modelos obtenidos fueron refinados mediante REFINE3D (<http://sysbio.rnet.missouri.edu/3Drefine/>) [14] y ModRefiner (<http://zhanglab.ccmb.med.umich.edu/ModRefiner/>) [15].

Los cuales mejoraron significativamente la calidad estructural del modelo para maximizar la similitud con las proteínas del alineamiento y minimizar errores de interacción.

D. Validación

Para validar las estructuras obtenidas en el paso anterior se evaluaron mediante programas como QMEAN [16] y ProSA-web [17]. Estos programas permiten analizar el modelo a través de comparaciones con estructuras cristalográficas en la base de datos Protein Data Bank (PDB).

III. RESULTADOS

Con el fin de encontrar proteínas con similitud estructural a la proteína de interés GNL3L, se realizó PSI-BLAST, BLASTp y Protein BLAST+ con la base de datos Protein Data Bank (PDB). Se encontró un dominio conservado entre GNL3L y una familia de proteínas de unión a GTP en bacterias; específicamente la proteína YlqF, la cual se encuentra involucrada en la biogénesis ribosomal. En la Tabla 1 se observan las dos proteínas que arrojaron el mejor resultado en la búsqueda.

Proteína	Max score	Total score	Query cover	E value	Ident	Accession
Chain A, Gnp-Bound Ylqf from T. Maritima	86.7	86.7	48%	2e-18	26%	3CNL_A
Chain A, Ylqf Gtpase from B. Subtilis	79.7	79.7	47%	4e-16	28%	1PUJ_A

Tabla 1. Proteínas con alineamientos significativos a GNL3L.

La estructura secundaria de estas dos proteínas es mayormente helicoidal. La proteína con código de acceso 3CNL está compuesta de un 35% de hélices alfa y 12% láminas beta; mientras que la proteína 1PUJ tiene un 42% de hélices alfa y 11% láminas beta.

Estas proteínas fueron alineadas con GNL3L usando la herramienta MUSCLE. Se encontró un porcentaje de identidad con GNL3L de 26% y 28% para 3CNL_A y 1PUJ_A respectivamente. Este porcentaje de identidad está entre los amino ácidos 126-398 de la secuencia de GNL3L, que corresponden al dominio GTPase.

Para la predicción de la estructura secundaria, la secuencia de aminoácidos fue ingresada en diversos programas de

predicción de estructura secundaria, como CFSSP, GOR, SOPMA, PHYRE2, y PSIPRED. En la Tabla 2 se resume el porcentaje de hélices alfa (H), coils (C), o extended strands (E) en GNL3L que se predice por cada herramienta. Los programas CFSSP y PHYRE2 presentan porcentajes de ambigüedad altos (¿representado por el símbolo?), que es la probabilidad de predecir tanto hélice alfa como extended strand para el mismo amino ácido, por lo que se descartan como herramientas confiables. Las demás herramientas (GOR, SOPMA, PSIPRED) en promedio predicen un 52% de hélices alfa y 10% extended strand, lo cual permite inferir que la proteína es principalmente helicoidal. Estos resultados están en concordancia con las estructuras secundarias de las proteínas que se encontraron son estructuralmente similares a GNL3L: 3CNL y 1PUJ

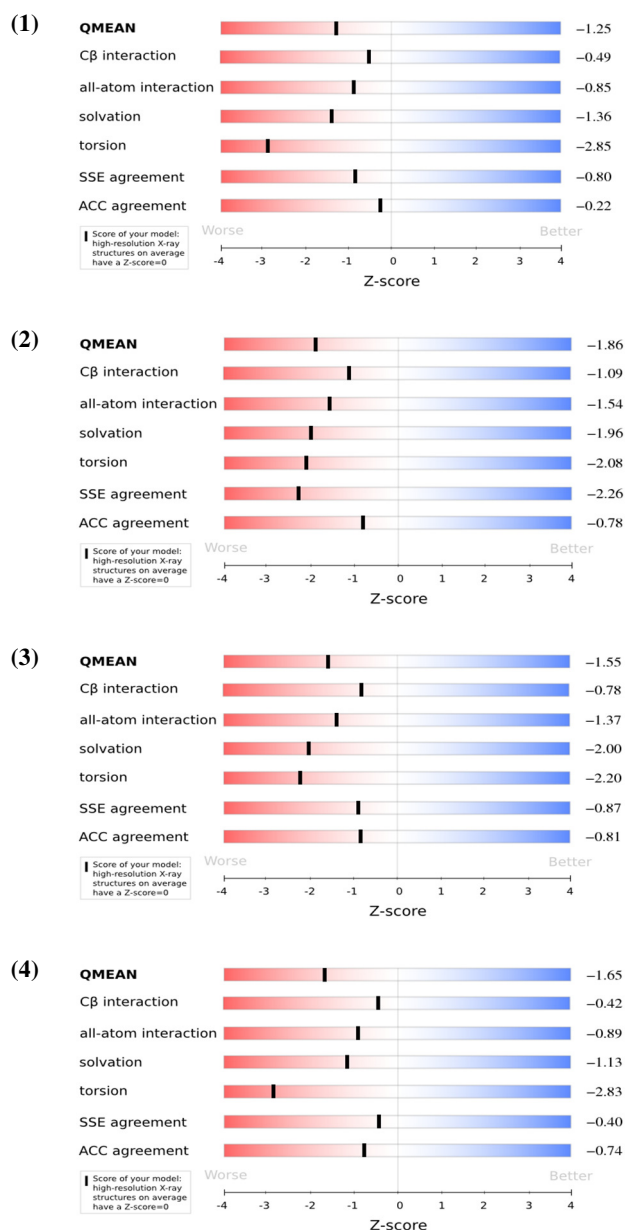
	H	E	C	?
CFSSP	77%	62%	----	11%
PHYRE2	54%	7%	35%	35%
GOR	46%	10%	43%	0%
SOPMA	52%	8%	32%	0%
PSIPRED	59%	12%	----	0%

Tabla 2. Predicción de estructura secundaria de GNL3L usando diferentes programas de predicción. Hélices alfa (H), extended strands (E), coils (C), y (?) predicción ambigua.

Para la predicción de la estructura terciaria de GNL3L se empleó el programa MODELLER. Este software, a partir de alineamientos óptimos obtenidos del PSI-BLAST, BLASTp y Protein BLAST+ permite predecir la estructura de la proteína objetivo. MODELLER posee una ventaja muy característica y es que permite crear modelos a partir de más de una secuencia molde de aminoácidos. Se obtuvieron cinco modelos a partir de alineamientos simples con 3CNL y 1PUJ, además de múltiples alineamientos usando las dos proteínas en conjunto.

En la predicción de estructura terciaria se producen un número considerable de modelos. La herramienta QMEAN (Qualitative Model Energy Analysis) permite calificar los principales aspectos geométricos de estos modelos, con el fin de escoger el mejor. El valor óptimo para la validación de un modelo es 0, aceptando desviaciones de + o - 1. En la Figura 2 se observa la evaluación por QMEAN de los cinco modelos obtenidos por MODELLER. El modelo 1 fue obtenido mediante alineamiento simple con 1PUJ_A, y el modelo 2 mediante alineamiento simple con 3CNL_A. Los modelos 3-5 fueron obtenidos mediante alineamientos múltiples con 1PUJ_A y 3CNL_A. Los mejores valores de QMEAN fueron arrojados por los modelos 1 (-1.25) y 5 (-1.24), a pesar de no estar en el rango óptimo (-1 a 1). Al comparar los evaluadores geométricos de estos modelos (C β interaction, all-atom interaction, solvation, torsión, SSE agreement, ACC agreement) la mejor evaluación la obtuvo el modelo 5, donde los valores de C β interaction, SSE agreement y ACC agreement se acercan a 0.

El modelo 5 fue refinado mediante la herramienta 3Drefine, que optimiza los enlaces de hidrógeno del modelo y minimiza la energía a nivel atómico [18]. En la Figura 2 se observa este modelo refinado como el modelo 6. En este modelo el valor de QMEAN se encuentra en el rango aceptable, y los valores de C β interaction, all-atom interaction, solvation, y ACC agreement se acercan a 0.



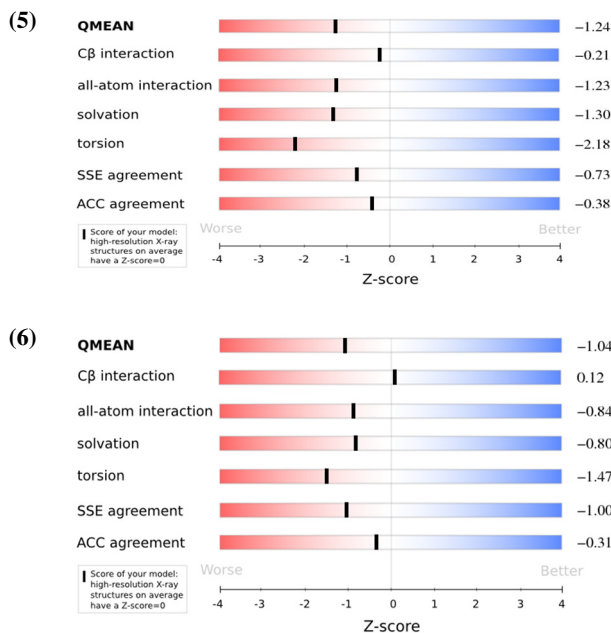


Figura 2. Evaluación por QMEAN de los principales aspectos geométricos de los modelos de estructura terciaria obtenidos para GNL3L. Modelos 1-5 obtenidos por MODELLER. Modelo 6 es la optimización del modelo 5 mediante 3Drefine.

Se realizó la validación del modelo 6 con la herramienta ProSA-web. El puntaje de calidad del modelo se grafica en un diagrama mostrando su relación con las puntuaciones de validación de todas las proteínas de tamaño similar y cuyas estructuras están determinadas experimentalmente y reposan en Protein Data Bank (PDB). Para el modelo 6 se obtuvo un z-score de -7.04 . Como se observa en la Figura 3(a) este resultado (representado por un punto negro) se encuentra dentro del rango óptimo de validación, además de ser similar a los resultados obtenidos con las proteínas molde: 1PUJ_A (-7.55) y 3CNL_A (-7.88).

En la Figura 3(b) se grafica la calidad del modelo en función de la energía a lo largo de la secuencia de aminoácidos. En general los valores que sobresalen la línea de horizonte (valores positivos) corresponden a partes problemáticas o erróneas del modelo. Como se puede ver, los residuos del modelo 6 presentan una baja tendencia a estar desordenados.

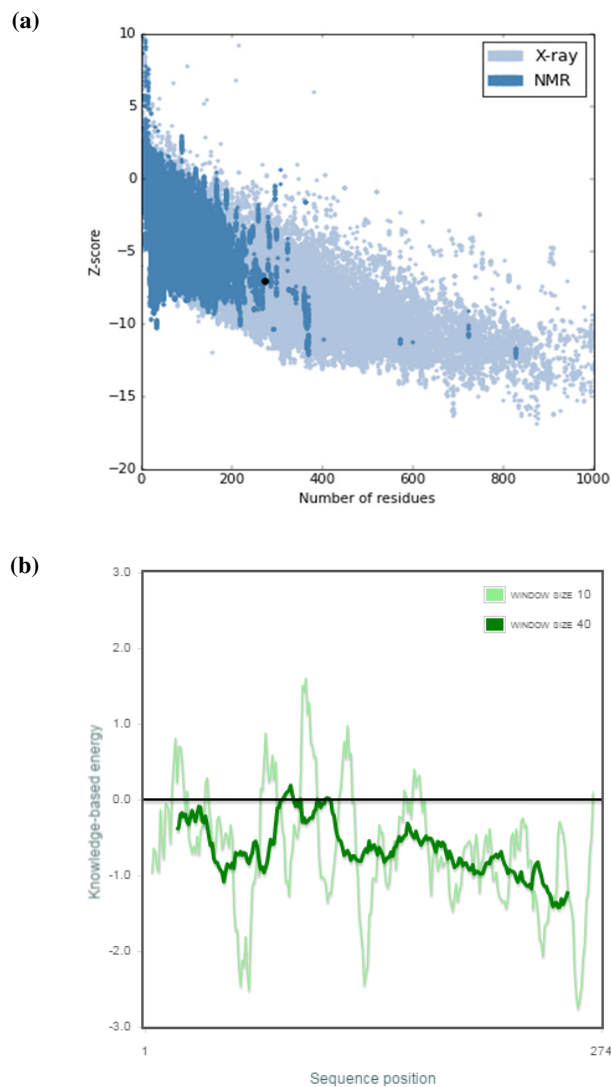


Figura 3. Resultados de validación del modelo 6 con ProSA-web. **(a)** El punto negro representa la puntuación del modelo 6 comparado con los resultados de otras proteínas de tamaño similar cuyas estructuras han sido determinadas experimentalmente por cristalografía de rayos X (azul claro) o por resonancia magnética nuclear (azul oscuro). **(b)** Grafico de energía de los aminoácidos de la secuencia. Los residuos por debajo de la línea representan la poca tendencia a estar desordenados.

Finalmente se visualizó la estructura 3D del modelo 6 de GNL3L usando el programa Jmol (ver Figura 4). La estructura presenta un gradiente de color donde las zonas rojas significan zonas poco fiables y las azules las más confiables. Tal como se esperaba, predominan las estructuras helicoidales (54%), y tiene un porcentaje reducido de estructuras laminares (8%). Esto es concordante con la estructura de las proteínas homólogas a GNL3L (3CNL y 1PUJ), y a la predicción de la estructura secundaria obtenida por GOR, SOPMA, y PSIPRED.

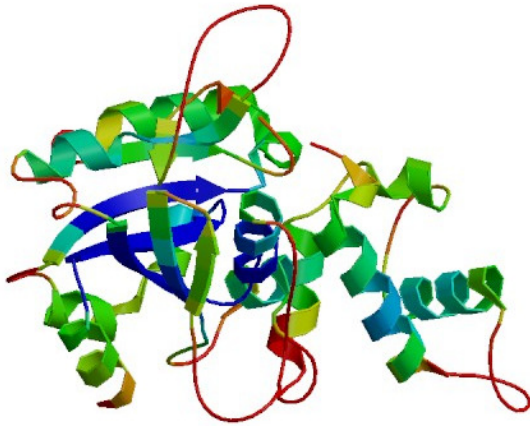


Figura 4. Potencial estructura 3D de GNL3L. Obtenida a partir de la predicción del modelo 6.

IV. CONCLUSIONES

De acuerdo con la predicción de la estructura secundaria se puede concluir que GNL3L es una proteína principalmente helicoidal cuya composición es de 52% hélices alfa y 9% laminas beta. Esta predicción es similar a la estructura secundaria de sus proteínas homólogas (3CNL y 1PUJ) a pesar de que el porcentaje de identidad entre ellas es bajo (26 – 28%) y solo corresponde a la sección de GNL3L entre los amino ácidos 126-398. Por lo tanto, las herramientas computacionales GOR, SOPMA y PSIPRED, son adecuadas para la predicción de estructuras secundarias, con la ventaja que son libres, automatizadas y sencillas de utilizar.

La predicción de la estructura terciaria, al basarse en algoritmos de homología, se ve limitada ya que requiere de proteínas cuyas estructuras ya estén experimentalmente determinadas y que tengan un alto porcentaje de homología a la proteína *query*. GNL3L y la familia de proteínas a la que pertenece (Nucleosteminas), es un grupo proteico que no está muy bien caracterizado. En un estudio similar [1] *in silico* de esta proteína se realizó la predicción con el fragmento 132-393 a.a. usando solo como secuencia molde a 1PUJ_A. Nuestro modelo comprende la sección 123-398 a.a. y está construido mediante el alineamiento múltiple de 1PUJ_A y 3CNL_A. Si bien el modelo final obtenido no tiene un QMEAN ideal (entre 0 y +/-1) el resultado (QMEAN:-1.04) es aceptable, considerando la escases de proteínas homólogas ya caracterizadas para usar como molde. No obstante, este modelo puede ser objeto de una mejor optimización de parámetros ya que las herramientas de refinado usadas, como 3Drefine, son automatizadas y al realizar un refinamiento manual se podrían mejorar los resultados obtenidos.

REFERENCIAS

- [1] F. Nazmi, M. A. Moosavi, M. Rahmati, and M. Hoessinpour-Feizi, "Modeling and structural analysis of human Guanine nucleotide-binding protein-like 3,nucleostemin.," *Bioinformation*, vol. 11, no. 7, pp. 353 – 358, 2015.
- [2] H. Xin, D. Liu, and Z. Songyang, "The telosome/shelterin complex and its functions.," *Genome Biol.*, vol. 9, no. 9, p. 232, Jan. 2008.
- [3] R. Bertorelle, E. Rampazzo, S. Pucciarelli, D. Nitti, and A. De Rossi, "Telomeres, telomerase and colorectal cancer.," *World J. Gastroenterol.*, vol. 20, no. 8, pp. 1940–50, Feb. 2014.
- [4] J. A. Londoño-Vallejo and R. J. Wellinger, "Telomeres and telomerase dance to the rhythm of the cell cycle," *Trends in Biochemical Sciences*, vol. 37, no. 9. pp. 391–399, Sep-2012.
- [5] A. Queisser, S. Heeg, M. Thaler, A. von Werder, and O. G. Opitz, "Inhibition of telomerase induces alternative lengthening of telomeres during human esophageal carcinogenesis," *Cancer Genet.*, vol. 206, no. 11, pp. 374–386, Nov. 2013.
- [6] Q. Zhu, L. Meng, J. K. Hsu, T. Lin, J. Teishima, and R. Y. L. Tsai, "GNL3L stabilizes the TRF1 complex and promotes mitotic transition," *J. Cell Biol.*, vol. 185, no. 5, pp. 827–839, Jun. 2009.
- [7] P. Invernizzi, F. Bernuzzi, A. Lleo, V. Pozzoli, M. Bignotto, P. Zermiani, A. Crosignani, P. M. Battezzati, M. Zuin, M. Podda, and C. Raggi, "Telomere dysfunction in peripheral blood mononuclear cells from patients with primary biliary cirrhosis.," *Dig. Liver Dis.*, vol. 46, no. 4, pp. 363–8, Apr. 2014.
- [8] M. R. K. S. Rao, G. Kumari, D. Balasundaram, R. Sankaranarayanan, and S. Mahalingam, "A Novel Lysine-rich Domain and GTP Binding Motifs Regulate the Nucleolar Retention of Human Guanine Nucleotide Binding Protein, GNL3L," *J. Mol. Biol.*, vol. 364, no. 4, pp. 637–654, Dec. 2006.
- [9] R. Y. L. Tsai and L. Meng, "Nucleostemin: A latecomer with new tricks," *International Journal of Biochemistry and Cell Biology*, vol. 41, no. 11. pp. 2122–2124, Nov-2009.
- [10] L. Meng, J. K. Hsu, Q. Zhu, T. Lin, and R. Y. L. Tsai, "Nucleostemin inhibits TRF1 dimerization and shortens its dynamic association with the telomere," *J. Cell Sci.*, vol. 124, no. 21, pp. 3706–3714, Nov. 2011.
- [11] P. B. Essers, T. C. Pereboom, Y. J. Goos, J. T. Paridaen, and A. W. MacInnes, "A comparative study of nucleostemin family members in zebrafish reveals

- specific roles in ribosome biogenesis,” *Dev. Biol.*, vol. 385, no. 2, pp. 304–315, Jan. 2014.
- [12] R. C. Edgar, “MUSCLE: a multiple sequence alignment method with reduced time and space complexity.,” *BMC Bioinformatics*, vol. 5, p. 113, 2004.
- [13] R. Sánchez and A. Sali, “Evaluation of comparative protein structure modeling by MODELLER-3.,” *Proteins*, vol. Suppl 1, pp. 50–58, 1997.
- [14] D. Bhattacharya and J. Cheng, “i3Drefine Software for Protein 3D Structure Refinement and Its Assessment in CASP10.,” *PLoS One*, vol. 8, no. 7, 2013.
- [15] D. Xu and Y. Zhang, “Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization,” *Biophys. J.*, vol. 101, no. 10, pp. 2525–2534, 2011.
- [16] P. Benkert, M. Künzli, and T. Schwede, “QMEAN server for protein model quality estimation.,” *Nucleic Acids Res.*, vol. 37, no. SUPPL. 2, pp. 510–514, 2009.
- [17] M. Wiederstein and M. J. Sippl, “ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins,” *Nucleic Acids Res.*, vol. 35, no. SUPPL.2, pp. 407–410, 2007.
- [18] D. Bhattacharya and J. Cheng, “3Drefine: consistent protein structure refinement by optimizing hydrogen bonding network and atomic-level energy minimization.,” *Proteins*, vol. 81, no. 1, pp. 119–31, Jan. 2013.