UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS
INSTITUTO DE INVESTIGACIÓN EN CIENCIAS BÁSICAS Y APLICADAS

CENTRO DE INVESTIGACIÓN EN DINÁMICA CELULAR

"Estructura residual en la región intrínsecamente desordenada de Escargot"

TESIS

QUE PARA OBTENER EL GRADO DE

***DOCTOR* EN CIENCIAS**

PRESENTA

Teresa Hernández Segura

DIRECTOR DE TESIS
Dra. Carmen Nina Pastor Colón

CUERNAVACA, MORELOS Enero 2021

Jurado Evaluador

Dr. Carlos Daniel Amero Tello

CIDC-UAEM

Presidente

Dr. Rodrigo Said Razo Hernández

CIDC-UAEM

Secretario

Dr. Enrique Rudiño Piñera

IBT-UNAM

Vocal

Dra. Verónica Mercedes Narváez Padilla

CIDC-UAEM

Vocal

Dr. Laura Domínguez Dueñas

FQ-UNAM

Vocal

Dr. Cesar Millán Pacheco

FF-UAEM

Suplente

Dra. Carmen Nina Pastor Colón

CIDC-UAEM

Suplente

Artículo de investigación asociado a la presente tesis

Hernández-Segura T, Pastor N.

"Identification of an α-MoRF in the Intrinsically Disordered Region of the Escargot Transcription Factor."

ACS Omega **5(29)**:18331-18341 doi: 10.1021/acsomega.0c02051 (2020)

Se encuentra anexado al final de la tesis.

Resumen

Las proteínas y regiones intrínsecamente desordenadas (IDPs e IDRs por sus siglas en inglés, respectivamente) no adoptan una estructura tridimensional bien definida en condiciones fisiológicas. Estas proteínas participan en muchas funciones biológicas y algunas de ellas están asociadas a diferentes enfermedades. Pequeñas regiones conocidas como MoRFs (Molecular Recognition Features, en inglés) pueden ser encontradas en IDPs o IDRs. Los MoRFs son regiones desordenadas que se encuentran en una constante transición estructural de orden y desorden, lo cual hace difícil su estudio con herramientas experimentales, ya que la mayoría de estas proporcionan información promedio del ensamble de diferentes conformaciones. Sin embargo, herramientas computacionales como las simulaciones de dinámica molecular (MD) permiten estudiarlas y caracterizar su ensamble conformacional a nivel atómico. En este proyecto estudiamos la proteína Escargot (Esg) de *Drosophila melanogaster*, la cual es un factor de transcripción de la familia Snail, y regula múltiples funciones, entre ellas el desarrollo del sistema nervioso. Estructuralmente, el dominio C-terminal de Esg es una región conservada que presenta dedos de Zn^{2+} e interactúa con ácidos nucleicos, mientras que el dominio N-terminal es una IDR. Actualmente, hay poca información estructural del N-terminal de Esg; la única anotación funcional en esta región consiste en dos motivos (P-DLS-K) que interactúan con la proteína de unión a C-terminal (CtBP). Sin embargo, el N-terminal de Esg ha sido asociado a funciones como degradación de proteínas, donde los dedos de Zn^{2+} no son necesarios. En este trabajo, presentamos el primer estudio bioinformático estructural del N-terminal de Esg de *Drosophila melanogaster* y diferentes ortólogos. Los resultados muestran que comparten un extremo N-terminal divergente a nivel de secuencia, pero probablemente conservado a nivel de desorden y estructura. Las regiones predichas como ordenadas presentan características de probables MoRFs, los cuales pudieran ser importantes para ejercer sus funciones como factores de transcripción. Por otra parte, presentamos el primer estudio estructural computacional del N-terminal de Esg de *Drosophila melanogaster*, en el cual investigamos la presencia de MoRFs. Analizamos una región de ~45 aminoácidos con probabilidad a formar estructuras ordenadas. A través de 54 μs de simulaciones de dinámica molecular (MD) usando CHARMM36 y el modelo de solvente implícito Generalizado de Born (GBSA), caracterizamos el espacio conformacional de esta región y encontramos un MoRF de ~16 aminoácidos. El MoRF adopta estructura de α-hélice y presenta pocos contactos de largo

alcance. También evaluamos y simulamos algunas mutaciones puntuales durante 24 μs de MD para probar la estabilidad del α-MoRF. Nuestros resultados mostraron que las mutaciones no desestabilizan el MoRF y llevan a una ganancia modesta de estructura residual, la cual puede ser efecto del campo de fuerza y modelo de solvente usado, y/o de contactos de largo alcance. Considerando lo anterior, se realizaron simulaciones de MD en solvente explícito simulando únicamente la región del α-MoRF. Los resultados muestran una labilidad distinta para la hélice silvestre y dos mutantes desestabilizantes, los cuales correlacionan con datos experimentales obtenidos por dicroísmo circular (CD).

Abstract

Intrinsically disordered proteins (IDPs) and regions (IDRs) lack a stable, well-defined structure in physiological conditions. These proteins are involved in multiple biological functions and some of them are associated to different diseases. Short regions known as MoRFs (Molecular Recognition Features) can be found in IDPs or IDRs. MoRFs are disordered regions that are in a constant structural transition of order and disorder, which is difficult the study with experimental methods, because the results are usually ensemble-averaged over the interconverting conformations. However, computational approaches like molecular dynamics simulations (MD) allow the study and characterization of conformational ensembles of IDPs with atomic detailed information. In this project we studied the Escargot protein (Esg) of *Drosophila melanogaster*, which is a transcription factor of the Snail family that regulates multiple cellular functions, including the development of the nervous system. Structurally, the C-terminal domain of Esg is a conserved and ordered region that has zinc fingers and interacts with nucleic acids, while the N-terminal domain of Esg is an IDR. Actually, there is not much structural information of the N-terminal domain of Esg; the only functional annotation in this region consists of two (P-DLS-K) motifs that interact with the C-terminal binding protein (CtBP). However, the N-terminal domain of Esg has been associated to functions like protein degradation, where zinc fingers are not necessary. In this work, we present the first bioinformatics structural study of the N-terminal domain of Esg of *Drosophila melanogaster* and diferent orthologs. The results show that they share a divergent N-terminal domain at the sequence level, but are probably conserved at the disorder and structure level. The predicted regions as ordered present features of probable MoRFs, which could be important in the role as transcription factors. In addition, we show the first structural computational study of the N-terminal domain of Esg of *Drosophila melanogaster*, where we researched the presence of MoRFs. We analized a region of ~45 amino acids with probability to form ordered structures. Through of 54 μs of molecular dynamics (MD) simulations using CHARMM36 and Generalized Born Implicit Solvent (GBSA), we characterized the conformational landscape of this region and we found a MoRF of ~16 amino acids. The MoRF adopts the structure of an α-helix and has few long-range contacts. We also evaluate and simulate some point mutations during 24 μs of MD to probe the stability of the α-MoRF. Our results show that the mutations do not destabilize the MoRF and lead to modest gains of residual structure, which could result from the force field and solvent model used, and/or long-

range contacts. Considering this, simulations of MD were carried out with explicit solvent only of the α-MoRF region. The results show a different lability of the wild-type helix and two destabilizing mutants, which correlates with experimental data obtained by circular dicroism (CD).

Dedicatoria

A mi familia por ser el pilar fundamental en mi vida

Gracias por su cariño, apoyo y comprensión

A mis guardianes: Candy, Tyson, Daysi, Merida, Lucas e Ivanna.

Por llenarme de vida, amor incondicional, lealtad y gratitud.

“No somos lo que debemos ser, tampoco somos lo que vamos a ser, ni mucho menos somos lo que queremos ser, pero…gracias a Dios no somos lo que éramos antes…”

Anónimo

Agradecimientos

Esta Tesis es el resultado de poco más de 5 años, los cuales me han parecido un abrir y cerrar de ojos. Quisiera expresar mi profundo agradecimiento a quien fue mi pilar principal en este proyecto, a mi directora de Tesis: la Dra. Nina Pastor. Gracias por darme la oportunidad de pertenecer a su grupo y apoyarme de manera incondicional desde el comienzo. Recuerdo que apenas me vió una vez en la vida y ya me había prestado uno de sus libros. Gracias por creer y confiar en mí, por ser más que una jefa un ejemplo de vida, una líder y una mamá académica ejemplar. Le admiro muchísimo y le agradezco eternamente cada uno de sus consejos, tiempo dedicado, paciencia, las porras, por echarme una mano en todo lo que necesité y los esfuerzos sobremanera ante las situaciones difíciles que se nos presentaron, pero sobretodo por siempre estar. Gracias por ayudarme a identificar cuando es momento de ir más despacio y apreciar las cosas buenas de la vida. ¡Vaya que esta aventura ha sido una odisea de las más agradables que he tenido!

En el rincón más entrañable de los afectos quiero agradecer a mi familia, mis padres, mis hermanos, mis abuelos, mis tíos y mis primos, por ser piezas fundamentales en mi vida y apoyarme en todo momento. Por ser siempre un estímulo e impulso a nuevos horizontes, y ser los principales promotores de mi sueños, sin pensar en limitaciones.

A White, gracias por estar siempre conmigo, por motivarme y ayudarme a crecer como persona. Sin duda, eres la hermana mayor que nunca tuve y parte de este logro ha sido gracias a ti.

A mis amigos y compañeros del cubo. A los de siempre, Ángel y Mark. Gracias por compartir cinco años continuos, y ser el trío perfecto de complicidad y terapias psicológicas, por ser un apoyo y sobretodo por motivarme a ser valiente. Sin duda hemos compartido vivencias muy bonitas. A Dianita y Edgar, gracias por el tiempo compartido, por su amistad y mostrarme el lado Zen de la vida.

A César Millán, gracias por ser parte de mi formación académica, de principio a fin. Gracias por ser mi amigo y brindarme tu apoyo siempre.

A Rodri, gracias por cada uno de tus consejos y motivaciones que me brindaste cuando lo necesitaba. Gracias por tu apoyo en todo momento y por prestarme los cerebros para culminar este proyecto. Gracias por tu linda e incondicional amistad.

A mi comité tutoral académico, la Dra. Verónica Narvaéz y el Dr. Carlos Amero, gracias por compartir su conocimiento, sus sugerencias y por brindarme su apoyo siempre.

A mi jurado de examen, gracias por aceptar ser parte del comité evaluador, por su tiempo a las revisiones y sugerencias, y por el apoyo brindado.

Gracias al Laboratorio de Dinámica de Proteínas del CIDC-IICBA donde fue realizada esta tesis. A las instalaciones de supercómputo: Centro Nacional de Supercómputo (IPICYT, San Luis Potosí), Yoltla (UAMI) y Laboratorio Nacional del Sureste de México (LNS Puebla).

Finalmente, gracias al Consejo Nacional de Ciencia y Tecnología de México (CONACyT) por la beca de doctorado 559324.

¡A cada uno de ustedes y de corazón, MUCHAS GRACIAS!

Índice

Índice de Figuras

Índice de Tablas

1. Introducción

1.1 Proteínas (IDPs) y regiones (IDRs) intrínsecamente desordenadas: Un nuevo concepto en el universo de las proteínas

La estructura molecular es crucial para la estabilidad y actividad de las proteínas. Se ha establecido que la secuencia de aminoácidos determina la estructura de una proteína, la cual a su vez determina su función[1]. Además, se ha demostrado que el universo de las proteínas incluye estructuras ordenadas (estructuradas), parcialmente ordenadas y completamente desordenadas (Figura 1).

Figura 1. Diferentes niveles de orden y desorden de una proteína. De izquierda a derecha: proteína ordenada, extremos N-terminal y C-terminal desordenados, conector desordenado, asa desordenada, dominios desordenados; proteína desordenada con algunas regiones ordenadas, proteína completamente desordenada y proteína desordenada y extendida. Las regiones ordenadas y desordenadas se muestran en color gris y rojo, respectivamente (Figura de Habchi et al., 2014)

Las proteínas ordenadas se caracterizan por presentar una estructura tridimensional bien definida. Las proteínas que no adoptan una estructura tridimensional bien definida se conocen como proteínas intrínsecamente desordenadas (IDPs), mientras que hay proteínas que contienen dominios estructurados en combinación con regiones desordenadas (IDRs)[2,3] (Figura 1). Las IDPs e IDRs presentan un alto grado de flexibilidad, lo que les permite adoptar múltiples conformaciones y establecer una gran diversidad de unión con otras moléculas[4]. El término “desordenado” lo han utilizado los cristalógrafos de rayos X para referirse a las regiones de una estructura que no se pueden resolver por cristalografía de rayos X debido a que presentan fluctuaciones en la red cristalina, mientras que el término “intrínseco” se refiere a la estructura y estados desordenados que son codificados por la secuencia proteíca[5]. Se ha encontrado que este tipo de proteínas son muy abundantes en la naturaleza y son importantes en varios procesos celulares.

1.2 Funciones de las IDPs e IDRs

La literatura reporta que el 30% de las proteínas eucariontes presentan IDRs[6], mientras que alrededor del 25% de las proteínas codificadas en el genoma humano son IDPs y el 40% contiene una región IDR cuya longitud es de al menos 30 aminoácidos[7,8]. Las IDPs e IDRs adoptan múltiples y diferentes conformaciones y sus funciones podrían establecerse a través de una estructura desordenada (Figura 2, estados conformacionales A, B, C o D), de la conversión entre estados desordenados (Figura 2, estados conformacionales C y D) y de transiciones entre un estado desordenado a ordenado (Figura 2, estados conformacionales D y E) y viceversa[9]. Frecuentemente las IDPs funcionan como nodos centrales en las redes de interacción proteína-proteína, ya que pueden establecer un gran número de interacciones con múltiples blancos moleculares[1,4,10,11].

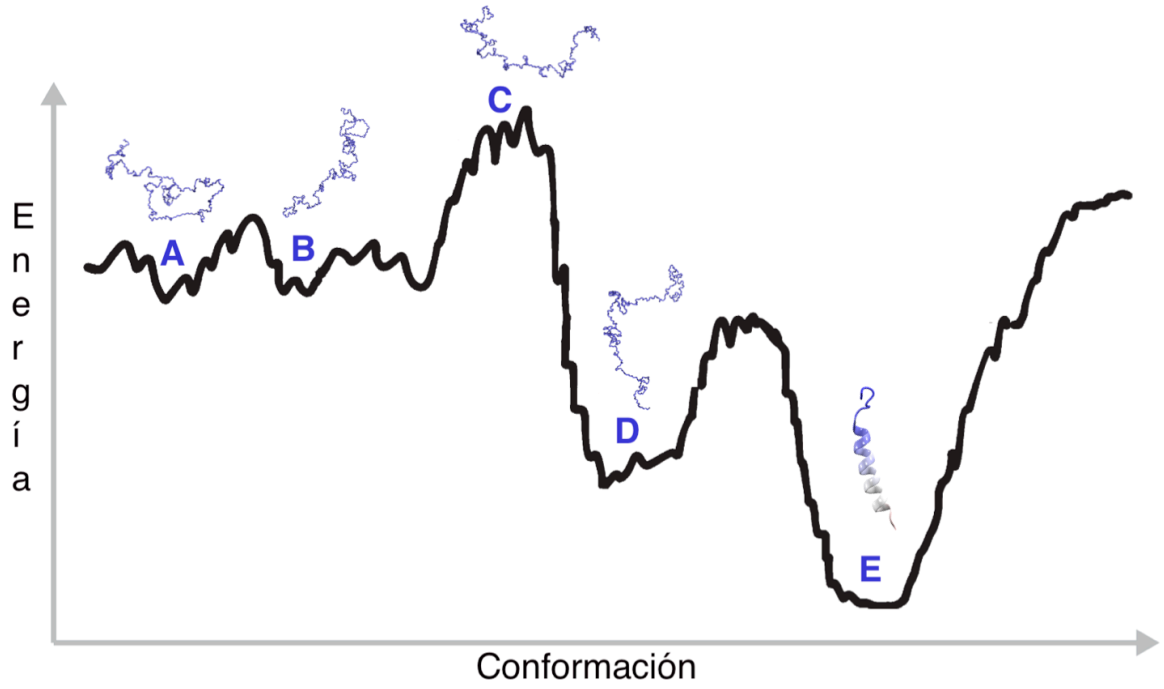


Figura 2. Esquema conformacional hipotético de una proteína IDP que muestra 5 conformaciones en un estado no unido, separadas por barreras energéticas. El paisaje conformacional de una IDP está representado por diferentes mínimos de energía, lo cual es debido a su alta flexibilidad estructural. En ausencia de su ligando, se pueden ver enriquecidos los estados conformacionales A, B y C. Sin embargo, una IDP puede llevar a cabo su función a través de la conversión entre dos conformaciones completamente desordenadas (estados conformacionales C y D) o bien, a través de la conversión de un estado desordenado (D) a ordenado (E), cuya

conformación se verá estabilizada una vez unida a su ligando. Las estructuras representativas fueron seleccionadas de Jensen M., y colaboradores (2014). Figura modificada de Perez A., y colaboradores (2017)[12].

Estas proteínas participan en diferentes funciones biológicas, tales como señalización celular, división celular, transporte intracelular, degradación de proteínas, regulación postranscripcional y control del ciclo celular[5,9,16,17]. También se han asociado a diferentes enfermedades como cáncer, enfermedades cardiovasculares y neurodegenerativas[7,10].

1.3 Mecanismos de unión de las IDPs e IDRs

La unión de las IDPs e IDRs con sus blancos moleculares puede ocurrir a través de los mecanismos de selección de conformación y/o ajuste inducido[13,14]. En el mecanismo de selección de conformación la IDP o IDR puede explorar conformaciones ordenadas y desordenadas en ausencia del blanco molecular, y algunas de esas conformaciones podrán interaccionar con el ligando dependiendo de su afinidad. En el caso del mecanismo de ajuste inducido, el cambio conformacional ocurre después de la unión con su blanco molecular, y este proceso es conocido como "plegamiento inducido por el molde" ("template folding"), donde la transición al estado plegado es definida por las interacciones con el ligando[15]. Estudiar los mecanismos de unión y las propiedades estructurales de las IDPs e IDRs permite conocer y proponer probables funciones y a su vez, entender los procesos fisiológicos y patológicos en los que participan.

1.4 Características de las IDPs e IDRs y herramientas de estudio

1.4.1 Características de las IDPs e IDRs a nivel de secuencia

La secuencia de aminoácidos de las IDPs es diferente a la de las proteínas estructuralmente ordenadas. Se ha sugerido que las IDPs están enriquecidas en aminoácidos polares (Arg, Gln, Ser, Glu y Lys) y aminoácidos que promueven el desorden (Gly y Pro), a diferencia de las proteínas ordenadas que son ricas en residuos hidrofóbicos (Ile, Leu, y Val), aromáticos (Trp, Tyr y Phe), los cuales en conjunto con los residuos Cys y Asn son considerados aminoácidos que promueven el orden [4,18,19].

Tomando en cuenta un conjunto de proteínas ordenadas y desordenadas, Uversky (2011) reportó que una baja hidrofobicidad y una carga neta alta son características de proteínas

IDPs, debido a que no proporcionan suficiente fuerza para la compactación estructural y contribuyen a la repulsión carga-carga en una proteína (Figura 3).

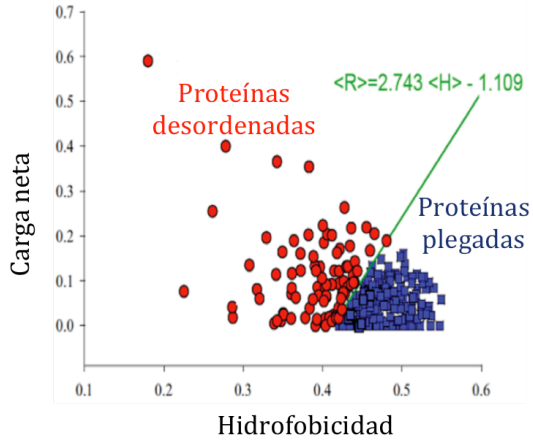


Figura 3. Gráfica de Hidrofobicidad vs Carga neta del promedio de un conjunto de proteínas desordenadas (círculos rojos) y ordenadas (círculos azules). Los dos conjuntos son separados por línea recta <R>=2.743<H>-1.109 en color verde (Figura modificada de Uversky, 2011).

A diferencia de las proteínas ordenadas, las IDRs e IDPs evolucionan rápido debido a que no presentan la restricción de mantener contactos terciarios que promuevan estructuras compactas y bien definidas. De acuerdo a su conservación evolutiva, se han identificado tres clases de IDRs: desorden flexible, desorden restringido y desorden no conservado[2]. Las IDRs que presentan desorden flexible mantienen conservado el desorden, pero no su secuencia. Las IDRs de desorden restringido se mantienen conservadas en desorden y en secuencia, mientras que las IDRs de desorden no conservado no se conservan ni en desorden ni en secuencia. Las IDRs de desorden flexible son las más comunes y se han asociado a funciones como señalización, regulación, reparación de DNA, ciclo celular, así como también pueden presentar modificaciones postraduccionales (PTMs) como glicosilación y fosforilación[11,20].

Además a las características anteriormente mencionadas, el hecho que una IDP o IDR evolucione rápidamente, provoca que haya motivos que se muevan de posición en la secuencia por procesos de deleción e inserción de aminoácidos. Siempre y cuando estos cambios de posición no afecten la función, es muy probable que el motivo pueda seguir conservándose con facilidad durante el proceso evolutivo[2]. Un ejemplo es el factor de transcripción p53, el cual participa en diferentes procesos como reparación de ADN, ciclo celular, apoptosis, autofagia y metabolismo. p53 presenta IDRs en su secuencia tales como el dominio de transactivación (TAD) presente en su N-terminal y el dominio C-terminal (CTD)[21,22], las cuales pueden adoptar diferentes conformaciones dependendo la

molécula con la cual se asocien. Así mismo, se considera que durante la evolución p53 presentó deleciones e inserciones de aminoácidos que provocaron el movimiento de posición de sus sitios de modificación SUMO (Small Ubiquitin-like Modifier, en inglés) antes y después del dominio de oligomerización observados en la secuencia de mosca y humano, respectivamente[2,23]. De acuerdo con lo anterior, alinear IDPs o IDRs entre proteínas ortólogas (proteínas homológas que se expresan en diferentes organismos y provienen de un mismo ancestro en común) para identificar motivos funcionales resulta una tarea difícil. Sin embargo, los análisis en la conservación y composición de secuencia han permitido saber que el desorden se puede conservar independientemente de la secuencia y a identificar preferencias funcionales[2]. Por ejemplo, IDRs con una alta conservación en residuos están presentes en proteínas involucradas en la regulación de la transcripción y unión a DNA, mientras que IDRs con una baja conservación en residuos en combinación con una alta conservación en la composición del tipo de aminoácidos se han asociado a funciones de ATPasa y nucleasa. Por su parte, IDRs que no muestran conservación en secuencia ni conservación en la composición de aminoácidos, son abundantes en proteínas de unión a iones metálicos[2].

1.4.2 Clasificación del desorden estructural presente en IDPs e IDRs

Las IDPs pueden adoptar diferentes niveles de desorden y ser clasificadas considerando como base la composición de aminoácidos. Este tipo de clasificación es relevante debido a que la composición de aminoácidos a menudo se muestra conservada entre ortólogos (organismos semejantes que provienen de un mismo ancestro en común) incluso si sus secuencias son poco conservadas[16].

Das y colaboradores (2015) han establecido un diagrama de estados que representa las diferentes clases conformacionales que pueden adoptar las IDPs con base en su composición (Figura 4). La clasificación de los estados depende de la fracción de los residuos cargados (FCR), la cual cuantifica la suma de las fracciones de los residuos cargados positivamente (f_+) y negativamente (f_-), y de la carga neta de residuos (NCPR). Las cuatro clases conformacionales corresponden a las regiones R1, R2, R3 y R4.

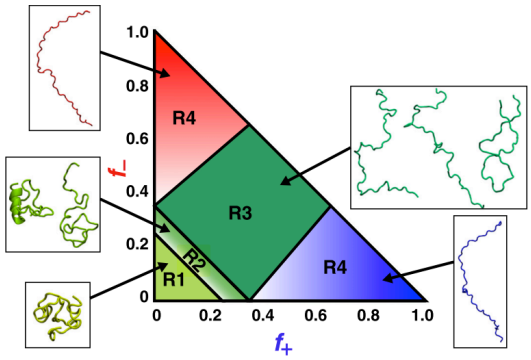


Figura 4. Diagrama de estados que considera las clases conformacionales que pueden adoptar las IDPs con base en su composición. (Figura de Das et al, 2015).

Las secuencias que caen en la región R1 en promedio adoptan conformaciones compactas y esféricas, de tipo glóbulo fundido, el cual corresponde a un estado de conformación compacta que puede contener elementos de estructura secundaria y es similar a un plegamiento terciario sin adoptar estructura tridimensional bien definida[5]. Las secuencias que caen en la región R2 por lo general presentan un comportamiento conformacional intermedio, es decir, pueden adoptar conformaciones tanto de glóbulo fundido como extendidas. La región R3 engloba secuencias que pueden adoptar conformaciones tipo asa u horquilla (hairpin, en inglés), mientras que las secuencias de la región R4 pueden adoptar conformaciones semiflexibles y extendidas. Se ha establecido que las regiones R1, R2 y R3 comprenden el 95% de las secuencias de IDPs conocidas[16].

1.4.3 Regiones MoRFs: Menos desordenadas y más ordenadas de lo esperado

El término estructura residual se relaciona con las conformaciones que establecen interacciones locales y de largo alcance[24], las cuales ayudan a establecer estructura secundaria. Una de las características que tienen las IDPs e IDRs es la presencia de elementos de estructura residual importantes para el reconocimiento molecular[25], mejor conocidos como MoRFs (Molecular Recognition Features, en inglés), los cuales son pequeñas regiones que consisten entre 10 a 70 aminoácidos, se someten a una transición de desorden a orden para adoptar estructura local y son estabilizadas cuando se unen a su blanco[11] (Figura 5). Este concepto lo introdujeron Fuxreiter y colaboradores en el 2004, describiendo a los MoRFs como elementos estructurales preformados en un estado libre o no unido, reconocidos por sus ligandos y presentes en IDPs e IDRs.

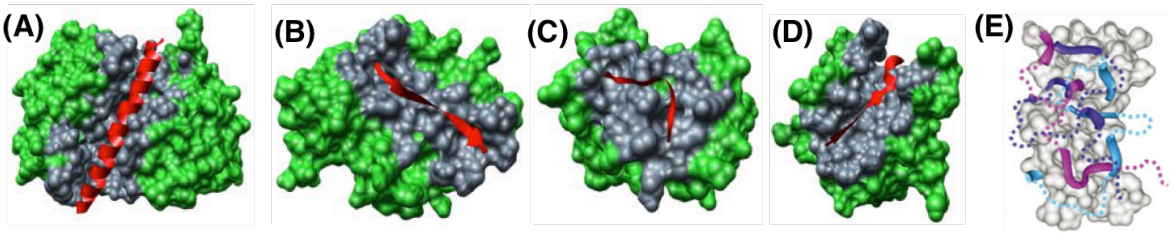


Figura 5. Ejemplos de MoRFs y Complejo difuso. A) α-MoRF, B) β-MoRF, C) ι-MoRF, D) MoRF complejo y E) Complejo difuso ("fuzzy"). La superficie verde con gris representa el blanco molecular, en rojo se representan los distintos tipos de MoRFs y en líneas punteadas azules y rosa se representa el desorden que presenta la IDP unida a su blanco molecular (Figura de Habchi et al., 2014).

Los MoRFs son clasificados de acuerdo a las conformaciones que adoptan, por ejemplo los α-MoRF adoptan estructuras de α-hélice (Figura 5A), β-MoRF adoptan estructuras β-plegada (Figura 5B), ι-MoRF adoptan conformaciones irregulares (conformaciones tipo asa, Figura 5C) y los MoRFs complejos (Figura 5D) adoptan conformaciones que resultan de alguna combinación de las anteriormente mencionadas[4]. Sin embargo, también hay reportes que las proteínas IDPs establecen complejos "difusos" ("fuzzy"), los cuales son aquellos complejos en donde la proteína IDP no se encuentra completamente plegada al unirse a su blanco (Figura 5E), manteniendo una parte de sí desordenada[1,26]. En general, la presencia de estructura residual les permite interaccionar con diferentes moléculas y participar en varios procesos biológicos.

1.4.4 Caracterización biofísica de las IDPs e IDRs

Las IDPs e IDRs presentan una gran variedad de movimientos, los cuales les permiten tener varios mínimos de energía que corresponden a diversos estados conformacionales dentro de un estado nativo[7]. Cada conformación puede estar asociada a diferentes funciones. A la fecha, se han determinado una gran variedad de complejos moleculares que involucran a IDPs o IDRs unidas a sus ligandos y reportadas en la base de datos de Protein Data Bank (PDB)[27]. La caracterización de los complejos y/o ensambles de los estados conformacionales que puede adoptar una IDP o IDR se han obtenido a través de diferentes técnicas a nivel experimental y computacional, entre las que se encuentran la Resonancia Magnética Nuclear (RMN), degradación proteica, dicroísmo circular (CD), dispersión de rayos X a bajo ángulo (SAXS) y simulaciones de dinámica molecular (MD)[4] y Monte Carlo (MC). Los estudios experimentales son útiles

para obtener información del promedio del conjunto de estructuras que pueden adoptar las IDPs e IDRs[21,28–30].

Gran parte de los estudios de IDPs combinan datos experimentales y computacionales. En las simulaciones de MD es necesario utilizar modelos estructurales para generar trayectorias (movimientos) y proponer conformaciones que describan los posibles estados nativos de IDPs e IDRs; estos resultados ayudan a confirmar y aclarar muchas de las observaciones obtenidas a nivel experimental. De esta manera, los métodos computacionales pueden ser de gran utilidad para complementar datos experimentales y caracterizar el espacio conformacional a nivel atómico.

1.4.5 Herramientas computacionales en el estudio de IDPs

Las simulaciones de MD nos permiten estudiar las características moleculares y movimientos de IDPs e IDRs. Sin embargo, la caracterización estructural por simulaciones de MD tiene como limitantes el tiempo requerido para muestrear de manera adecuada las conformaciones y la precisión de los parámetros de los campos de fuerza, así como los modelos de solvente (explícito e implícito) utilizados. Los campos de fuerza fueron desarrollados y parametrizados principalmente para proteínas plegadas[28,31] y han sido bien evaluados, pero su aplicación a IDPs requiere de continua atención y mejora[21]; variaciones en los campos de fuerza y uso de diferentes protocolos con modelos de solvente explícito e implícito han sido aplicados a diferentes sistemas de IDPs e IDRs (Tabla 1).

Los movimientos de los cambios conformacionales en IDPs ocurren en varios órdenes de magnitud (desde ps hastas ms)[8], y esto implica un costo computacional, considerando el tamaño del sistema[32]. Los modelos de solvente explícito usan la inclusión de moléculas de agua de manera explícita, y este modelo de solvente resulta insuficiente para simular escalas de tiempo largos o sistemas grandes[33] por el alto costo computacional. En este caso, los modelos de solvente implícito son una opción útil porque solo consideran de manera explícita las coordenadas atómicas del soluto; las moléculas de agua son representadas en un medio continuo infinito con propiedades del agua, y el tamaño del sistema es reducido con un costo computacional menor. En consecuencia, con este modelo de solvente es posible tener un balance de precisión y velocidad[33,34]. Sin embargo, se ha reportado que variaciones en los campos de fuerza y modelos de

solvente afectan los resultados de las simulaciones de IDPs e IDRs. Los modelos de solvente implícito generan conformaciones más estructuradas y compactas[32]. En el caso de modelos de solvente explícito, los modelos de agua son importantes para tener precisión en las simulaciones de IDPs, ya que algunos de ellos pueden llevar a una sobrecompactación estructural[28,31] o a una mayor correlación con datos experimentales en la caracterización de los ensambles conformacionales[31]. Además, algunos campos de fuerza generan una mayor compactación que otros, así como también una sobre-estabilización estructural[31]. Actualmente, no existe un protocolo específico para simular IDPs e IDRs, pero existe una búsqueda continua para desarrollar y mejorar los campos de fuerza y puedan aplicarse tanto a proteínas plegadas como desordenadas.

Tabla 1. Protocolos de simulaciones de MD de IDPs e IDRs en los últimos 5 años.

Sistema	Solvente	Campo de fuerza	Tiempo total (acumulado)	Referencia
Bcl-2 (239 residuos)	Explícito (TIP3P)	CHARMM27	25 ns	35
Región rFLD (residuos 60-77)			1.0 μs	
Caja 3 del motivo CoRNR de NCOR1 (residuos 2258 – 2277)	Explícito (TIP3P)	CHARMM22*	8 µs	36
COR15A (89 residuos)	Explícito (Glicerol-TIP4P)	OPLS-AA	300 ns para cada sistema	37
COR15B (90 residuos)				
α-sinucleína (140 residuos)	Implícito (GBSA)	CHARMM27	10 ns para cada sistema	38
Dominio de translocación de Colicina N (90 residuos)				
Dominio K18 de Tau (130 residuos)			30 ns	
α-sinucleína (residuos 42 - 63)	Explícito (SPC)	GROMOS 54A7	14 µs	29
C-Myb (residuos 291-315)	Explícito (TIP3P, TIP4P-EW, TIP5P)	ff99IDPs ff99SBildn	1.9 μs (TIP3P) 950 ns (TIP4P-EW) 950 ns (TIP5P)	39
			1 μs (TIP3P) 500 ns (TIP4P-Ew) 500 ns (TIP5P)	
$A\beta_{40}$ (30 residuos)	Explícito (TIP3P, TIP3Pm)	CHARMM36	3.2 μs (TIP3P) 3.2 μs (TIP3Pm)	40
		CHARMM22	3.2 μs (TIP3P)	
		CHARMM22*	3.2 μs (TIP3Pm)	
		OPLS-AA	3.2 μs (TIP3Pm)	

Péptido desplegado: PepG (9 residuos)	Explícito (TIP4P)	ff03w	14.04 μs	41
Péptido desplegado: PepW (9 residuos)			4.04 μs	
Dominio CTD de p53 (residuos 375 – 388)			9.64 μs	
$A\beta_{42}$ (42 residuos)	Explícito (TIP4P-Ew)	ff99SB*ILDN ff99SBILDN-NMR ff99SB CHARMM22*	6.4 μs para cada campo de fuerza	42
	(TIP3P)	OPLS ff99SB		
RAM-Región desordenada del receptor Notch (WT and 13 variantes) (residuos 1-100)	Implícito (ABSINTH)	OPLS-AA	9.20×10^8 pasos MMC para cada sistema	43
C-terminal de la γ-tubulina (35 residuos) (WT and 3 variantes)	Implícito (ABSINTH)	OPLS-AA	1.20×10^8 pasos MMC para cada sistema	44
9 péptidos pequeños (9 residuos)	Explícito (TIP3P)	ff14SB ff14IDPSFF	10 μs para cada sistema	45
Proteína HIV-1 Rev (23 residuos)			1 μs	
$A\beta_{40}$ (40 residuos)	Explícito (TIP4P)	ff99SBws	743.760 ns ($A\beta40$) 740.715 ns ($A\beta42$)	46
$A\beta_{42}$ (42 residuos)		ff03ws	750 ns para cada sistema	
p53 TAD2 (residuos 41-62)	Explícito (TIP3P)	ff03 CHARMM27 OPLS-AA/L ff99SB-ILDN CHARMM36m	3 µs para cada campo de fuerza	21
Syt-1IDR (residuos 80-141)	Implícito (GBSA)	CHARMM36	3.180 μs	47
Región Core de la IDR (residuos 97-130)			4.18 μs	
$A\beta_{1-40}$ (WT y una variante)	Implícito (GBSA)	CHARMM36	1.098 μs para cada sistema	48
SH4UD (95 residuos)	Explícito (TIP3P)	ff03ws	10.20 μs	49
Proteína IN (55 residuos) $A\beta_{1-42}$ (42 residuos) Cola de la histona H4 (26 residuos)	Explícito (TIP3P, OPC)	ff99SB	Para cada modelo de agua: 15 μs (Proteína IN) 10 μs ($A\beta_{1-42}$ y cola de la histona H4)	28
	Implícito (GB Neck)		1 μs para cada sistema	
N-terminal del α-MoRE	Explícito	ff99SB	100 μs	50

(residuos 484 to 504)	(TIP4P-D)			
4E-BP2 (WT y dos variantes) (residuos 18 a 62)	Explícito (TIP3P)	CHARMM36m ff99SB-ILDN	19.20 μs para cada sistema y campo de fuerza	[51]

2. Antecedentes

2.1 Familia Snail: Factores de transcripción con abundantes IDRs

Actualmente se sabe que proteínas relacionadas a funciones como la transducción de señales y transcripción, están enriquecidas de IDRs. Por ejemplo, el 82.63% al 94.13% de los factores de transcripción presentan IDRs, las cuales varían en longitud y composición[52,53]. Snail es una superfamilia de factores de transcripción que puede ser dividida en dos familias independientes: Snail y Scratch[54]. Cuatro genes de la familia Snail son expresados en insectos (Snail, Escargot, Worniu y Scratch) y tres en vertebrados (Snail, Slug y Scratch). La familia Snail se caracteriza por presentar dedos de Zn^{2+} (ZNFs) en su extremo C-terminal, dominios que son necesarios en la regulación transcripcional y unión a DNA.

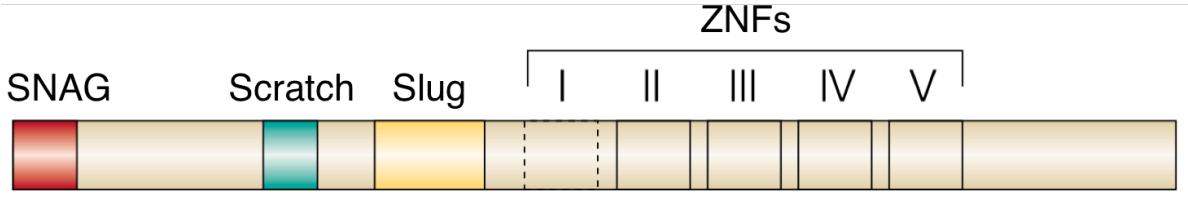


Figura 6. Dominios conservados y ZNFs presentes en los miembros de la superfamilia Snail. En color rojo se muestra el dominio SNAG (Snail/Gfi), en verde el dominio Scratch, en amarillo el dominio Slug y los ZNFs (del I al V). Figurada modificada de Nieto (2002).

Por otra parte, el extremo N-terminal de la familia Snail es divergente entre vertebrados e invertebrados, así como entre ellos mismos. Se ha reportado que Snail, Worniu y Escargot comparten un pequeño segmento conservado localizado en su N-terminal, conocido como NT box (CPLKKRP) cuya función no ha sido determinada, pero se ha sugerido que está involucrado en la localización nuclear[55]. En cuanto a los homólogos en vertebrados, éstos contienen dos motivos conservados en su extremo N-terminal, conocidos como dominio SNAG (Snail/Gfi) de siete aminoácidos (MPRSFLVK) y dominio Scratch. Ambos dominios actúan como un represores transcripcionales[56]; sin embargo el dominio Scratch está menos caracterizado que el dominio SNAG[57] (Figura 6). Así mismo,

se ha reportado que los genes Snail de *Drosophila melanogaster* (a excepción de Scratch) contienen dos dominios de unión a CtBP. De esta manera se ha sugerido que la actividad represora de las proteínas Snail se encuentra conservada a través de diferentes mecanismos: co-represión a través de CtBP o por el dominio SNAG.

Otra función en la que participa esta familia es la homeostasis celular, a través de la síntesis y degradación de proteínas. El sistema de ubiquitina-proteosoma (UPS) es esencial para múltiples procesos fisiológicos, entre ellos el cáncer, proliferación celular, apoptosis, y enfermedades neurodegenerativas, entre otras[58–60]. Se ha reportado que Esg y Slug interactúan con otras proteínas, como Daughterless, y promueven su eliminación en una vía UPS. Esta función no requiere ZNFs funcionales, por lo que probablemente sea una función adscrita al N-terminal no estructurado[61]. Una de las funciones más importantes de los factores de transcripción de la familia Snail es su participación en el desarrollo embrionario. Mutaciones en Slug han sido vinculadas al Síndrome de Waardenberg, cuyas principales características son sordera congénita, problemas visuales y mechones de cabello blanco[61].

2.2 El misterio estructural y funcional presente en el factor de transcripción Escargot (Esg)

Escargot (Esg) es un factor de transcripción de la familia Snail expresado en *Drosophila melanogaster (Dme)*. Esg está compuesto por 470 aminoácidos y contiene cinco ZNFs (cuatro de los cuales son de tipo C_2H_2 y uno de tipo CCHC) en su C-terminal[56,62,63]. En porcentajes de identidad de su extremo C-terminal, los cuatro genes de *Dme* (Snail, Escargot, Worniu y Scratch) comparten entre 76% a 85%, mientras que la identidad entre *Dme* y vertebrados es entre 50% a 70%. Estos porcentajes sugieren que el extremo C-terminal de Esg y sus homólogos se encuentra conservado tanto en estructura como en función[56,62]. En la búsqueda de proteínas que interactúan con Esg, se identificaron dos dominios (P-DLS-K) de unión a la proteína CtBP (Proteína de unión C-terminal de *Dme*) en su extremo N-terminal[63,64].

Se sabe que Esg es expresado en testículos, y otros tejidos y poblaciones celulares de *Dme*, como son neuroblastos, tracto digestivo, células madre y enterocitos. Entre las funciones que se le han vinculado a Esg están: desarrollo de histoblastos abdominales y sistema traqueal, ayudar a mantener la diploidía de los discos imaginales, diferenciación

neuronal, regulador del ciclo celular y comunicación celular, represor o activador de células madre intestinales y enteroblastos, y ayudar a mantener un proceso parcial de EMT (Transición de Epitelio a Mesénquima), entre otras[59,64–67].

Mutaciones en Esg producen pérdida de función como factor de transcripción y generan fenotipos letales en *Dme*[68], cuyas principales características son defectos en la cutícula, pérdida de cerdas y tergitos, además de mostrar anormalidades en las estructuras que derivan de discos imaginales como abdomen (no hay proliferación de histoblastos), alas y patas (más extendidas), cabeza y ojos (reducción de tamaño)[69]. Actualmente, la única mutación puntual reportada en Esg presente en el C-terminal es G387→E, la cual genera un fenotipo letal[69,70]. En ese mismo trabajo se habla de polimorfismos en el N-terminal, P149→Q y M196→I, que se consideran inocuos[70]. Estudiar el N-terminal de Esg y sus ortólogos es importante, porque además de ser su extremo N-terminal muy divergente, se ha reportado que los factores de transcripción de la familia Snail probablemente cambien su conformación al unirse a ADN o ARN[10,71,72].

3. Fundamento teórico

3.1 Justificación

La familia Snail es una familia de factores de transcripción. Su característica más conservada y la función más entendida a la fecha depende de la región de dedos de Zn^{2+} (ZNFs) los cuales se encuentran en su extremo C-terminal. Se ha reportado que los factores de transcripción de la familia Snail son muy divergentes en su extremo N-terminal, y el entendimiento de la estructura y función de esta región es aún limitado por la ausencia de modelos estructurales. A la fecha no hay un estudio bioinformático estructural que compare el extremo N-terminal de Esg de *Dme* y sus ortólogos. Realizar un análisis comparativo permitirá conocer regiones que se conservan a nivel de secuencia y estructura, las cuales pudieran ser importantes para realizar funciones similares o bien, participar como regiones compensatorias que están presentes en algunos genes y en otros no, pero que les permiten ejercer sus funciones como miembros de la misma familia.

Por otra parte, se ha reportado que los factores de transcripción presentan IDRs, las cuales son esenciales para poder llevar a cabo sus diferentes funciones, entre ellas el reconocimiento molecular. Son pocas las IDRs de factores de transcripción que han sido

caracterizadas estructuralmente, debido a que adoptan diferentes conformaciones que se encuentran en una constante transición estructural de orden y desorden, y en algunas ocasiones pueden adoptar estructura secundaria e incluso contactos terciarios. De hecho, se sabe que la presencia de α-MoRFs es abundante en factores de transcripción[53]. Considerando lo anterior y debido a que el N-terminal de Esg de *Dme* ha sido asociado a funciones como degración de proteínas (donde los ZNFs no son necesarios), realizar una búsqueda de MoRFs en esta región, ayudará a encontrar similitudes estructurales y comprender mejor la relación estructura-función de los mecanismos que regulan a este factor de transcripción, así como los procesos fisiológicos en los que participa.

3.2 Hipótesis

El extremo N-terminal no estructurado de Esg de *Dme* y ortólogos presenta características estructurales similares. A su vez, el N-terminal de Esg de *Dme* presenta MoRFs.

3.3 Objetivos

3.3.1 Objetivo general

Localizar en el N-terminal no estructurado de Esg de *Dme* características estructurales similares con sus ortólogos, así como la presencia de estructura residual que pudiera participar en procesos de reconocimiento molecular.

3.3.2 Objetivos particulares

1. Realizar un análisis del extremo N-terminal no estructurado de Esg de *Dme* y sus ortólogos, calculando perfiles de desorden y realizando un análisis a nivel de secuencia y composición.

2. Generar un ensamble de estructuras y un muestreo amplio de la región con probabilidad a ordenarse presente en el N-terminal de Esg de *Dme* a partir de simulaciones de dinámica molecular.

3. Identificar y caracterizar MoRFs del extremo N-terminal de Esg de *Dme,* calculando estructura secundaria, contactos y puentes de hidrógeno.

4. Proponer residuos clave para la estructura de los MoRFs del extremo N-terminal de Esg de *Dme*.

5. Comparar los resultados obtenidos de las simulaciones de Dinámica Molecular con datos experimentales obtenidos por Dicroísmo Circular.

4.- Estrategia computacional

4.1 Predicción de desorden a nivel de secuencia de Esg

Las secuencias de Esg de *Dme* y sus ortólogos (Tabla 2), al igual que sus homólogos (Tabla 3) fueron obtenidas de acuerdo a los resultados del blast[73] a nivel de proteína en formato FASTA. El análisis de desorden se desarrolló utilizando los predictores Metadisorder[74], MFDp2[75], AUCPred[76] y SPOT-disorder[77], los cuales predicen mejores resultados que el resto de los predictores de desorden reportados a la fecha[78]. Metadisorder es un metaservidor que genera un promedio de la salida de varios predictores de desorden, esto con el objetivo de mejorar la predicción a diferencia de usar un sólo predictor. Se utilizó la salida de Metadisorder (URL http://genesilico.pl/metadisorder/) de sus tres diferentes variantes: MetaDisorderMD, MetaDisorderMD2 (una variante de MetaDisorderMD con diferente función de calificación) y MetadisorderMD3 (basada en métodos de reconocimiento de plegado o "fold recognition"). MFDp2 (URL http://biomine.cs.vcu.edu/servers/MFDp2/) predice el desorden a nivel de secuencia y residuo. Por su parte AUCPred y SPOT-disorder pueden predecir IDRs pequeñas y grandes. AUCpreD (URL http://raptorx2.uchicago.edu/StructurePropertyPred/predict/) predice desorden considerando información a nivel de secuencia, evolutiva y estructural, mientras que SPOT-disorder (URL https://sparks-lab.org/server/spot-disorder-single/) considera información a nivel de secuencia. Los predictores utilizados asignan una calificación de desorden para cada aminoácido de la secuencia. Se calculó el porcentaje de desorden de la proteína Esg de *Dme* y sus ortólogos a través de un consenso de los resultados obtenidos por los predictores. Residuos con calificaciones por arriba de 0.5 fueron considerados como desordenados. El resultado se graficó utilizando Matplotlib 2.2.3 [79,80].

4.2 Análisis bioinformático del factor de transcripción de Esg

La secuencia de la proteína completa de Esg expresada en *Drosophila melanogaster* (*Dme*-Esg) (470 aminoácidos) fue obtenida de UniProt[81], (Uniprot ID:P25932, URL htttp://www.uniprot.org/) y se realizó un blast de proteínas utilizando el servidor de NCBI[82] (URL https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins). En el blast, se utilizó la secuencia de los ZNFs de *Dme*-Esg la cual sirvió de anzuelo para poder identificar proteínas ortólogas de distintos filos de metazoarios (Tabla 2 y Figura 7).

Tabla 2. Proteínas ortólogas de Esg pertenecientes a diferentes filos de Metazoarios.

Organismo	Abreviación
Amphimedon queenslandica	*Aqu*
Acropora millepora	*Ami*
Patella vulgata	*Pvu*
Platynereis dumerilii	*Pdu*
Fasciola hepatica	*Fhe*
Hypsibius dujardini	*Hdu*
Drosophila melanogaster	*Dme*
Strongylocentrotus purpuratus	*Spu*
Saccoglossus kowalevskii	*Sko*
Branchiostoma belcheri	*Bbe*
Ciona intestinalis	*Cin*
Danio rerio	*Dre*
Xenopus tropicalis	*Xtr*
Amazona aestiva	*Aae*
Ophiophagus hannah	*Oha*
Homo sapiens	*Hsa*

De acuerdo a los resultados obtenidos, se seleccionaron factores de transcripción de la familia Snail (Sna1 y Sna2), así como proteínas de *Amphimedon queenslandica* y *Branchiostoma belcheri*, las cuales no son proteínas identificadas de la familia Snail pero se consideraron al presentar ZNFs y al ser una región que alineó bien con el extremo C-terminal de *Dme*-Esg. Se realizó un árbol filogénetico de Esg utilizando la herramienta Taxonomy browser de NCBI (URL https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi). De esta manera, al tener una lista de proteínas ortólogas a Esg de distintos filos, aseguramos tener una muestra variada para estudiar la divergencia presente en el N-terminal de los factores de transcripción de la familia Snail. A su vez, se utilizó la secuencia completa de *Dme*-Esg y se realizó un blast en el cual no se consideró a la familia *Drosophilidae*; esto ayudó a identificar proteínas que comparten similitud en secuencia con *Dme*-Esg (Tabla 3, Figura 10) y las cuales corresponden a distintos géneros de moscas.

Por otra parte, se realizó un alineamiento de secuencias utilizando Clustal Omega[83]. Este alineamiento se generó para mapear las diferencias en secuencia presentes en el N-termina de Esg de *Dme* y sus ortólogos en moscas, y conocer su grado de conservación. También, se realizó un análisis de propiedades estructurales del extremo N-terminal de *Dme* y sus ortólogos utilizando los servidores CIDER[84] (Classification of Intrinsically Disordered Ensemble Relationships) e IDDomainSpotter[85]. CIDER es un servidor que genera información de las propiedades en composición de las proteínas IDPs tales como la fracción de aminoácidos positivos y negativos, los

cuales influyen directamente en su tamaño, forma y amplitud de las fluctuaciones conformacionales. Por su parte, IDDomainSpotter puede predecir dominios conservados presentes en IDRs e IDPs independientemente de su longitud. IDDomainSpotter genera diagramas que muestran la calificación de diferentes características de secuencia; por ejemplo incluye Phe, Tyr y Gly (+FYG), para identificar IDRs que están involucradas en procesos de separación de fases líquido – líquido, y Leu, Val e Ile (+LVI), ya que las IDRs no están enriquecidas en residuos hidrofóbicos. También incluye Arg y Lys (+RK) y carga neta definida por Arg, Lys, Asp y Glu (+RK-DE), para identificar dominios de unión a DNA presentes en IDPs, tales como factores de transcripción. Así mismo, identifica regiones hidrofílicas enriquecidas por aminoácidos que promueven el desorden (Pro, Ser y Thr) y deficientes en Arg y Lys (+PST-RK), las cuales se predicen como regiones altamente desordenadas. Distinguir este tipo de propiedades ayuda a describir a las IDPs e IDRs a nivel químico y estructural.

Tabla 3. Proteínas homólogas de Esg pertenecientes al filo Artrópodos

Organismo	Abreviación
Drosopholila melanogaster	*Dme*
Scaptodrosophila lebanonensis	*Sle*
Bactrocera dorsalis	*Bdo*
Zeugodacus cucurbitae	*Zcu*
Lucilia cuprina	*Lcu*
Ceratis capitata	*Cca*
Rhagoletis zephyria	*Rze*
Stomoxys calcitrans	*Sca*
Musca domestica	*Mdo*

4.3 Predicción de estructura secundaria de Esg

Se utilizó el predictor de estructura SPARKS-X[86] (URL https://sparks-lab.org/server/sparks-x/). para obtener modelos estructurales y se evaluó el perfil de estructura secundaria de las regiones con probabilidad a ordenarse presentes en el N-terminal de Esg de *Dme* y sus ortólogos utilizando CARMA[87]. Cada modelo fue previamente optimizado con CHARMM-GUI[88] (URL: http://www.charmm-gui.org): se les añadieron hidrógenos, se minimizaron 200 pasos usando el esquema de Steepest descent sobre los hidrógenos y 400 pasos sobre todos los átomos para eliminar choques estéricos.

4.4 Generación del ensamble estructural de la región S2 de *Dme*-Esg

Se utilizaron diferentes predictores de estructura para obtener estructuras iniciales de la región S2 presente en el N-terminal de *Dme*-Esg (residuos 111 al 155). Los predictores utilizados fueron HHPred[89] (URL https://toolkit.tuebingen.mpg.de/tools/hhpred), I-Tasser[90] (URL https://zhanglab.ccmb.med.umich.edu/I-TASSER/), Phyre2[91] (URL http://www.sbg.bio.ic.ac.uk/~phyre2/html/page.cgi?id=index), QUARK[92] (URL https://zhanglab.ccmb.med.umich.edu/QUARK/) y SPARKS-X[86] (URL https://sparks-lab.org/server/sparks-x/). Estos predictores utilizan métodos de modelado por homología, fold recognition y *ab initio*. Se obtuvieron 27 modelos estructurales como ensamble inicial, que fueron visualizados utilizando VMD 1.9.2[93] y se generaron los archivos de entrada para NAMD con CHARMM-GUI[88,94] (URL http://www.charmm-gui.org) para su posterior simulación.

4.5 Generación de mutantes de la región S2 de *Dme*-Esg

Para generar los modelos de las mutantes, se seleccionaron 4 modelos del ensamble inicial que presentaron diversidad estructural (HHP, PI, SP6 y SP9). Para cada uno de los modelos, se mutaron los residuos E141P, E141L, E141V, E141A_T145A, R144A, T145V y Q149P. Las mutantes se construyeron utilizando CHARMM-GUI y se generaron los archivos de entrada para NAMD para su posterior simulación.

4.6 Simulaciones de dinámica molecular de la región S2 de *Dme*-Esg con CHARMM36 y solvente implícito

Estas simulaciones de dinámica molecular se desarrollaron usando el programa NAMD 2.10 (URL http://www.ks.uiuc.edu/Research/namd/)[95], aplicando el campo de fuerza CHARMM36 (URL http://mackerell.umaryland.edu/charmm_ff.shtml)[96], usando como solvente implícito el modelo Generalizado de Born[97] GBSA (por sus siglas en inglés Generalized Born/Surface Area), y un paso de integración de 2 fs. Las simulaciones se realizaron a temperatura constante (298 K). Se aplicó una constante dieléctrica de 80, una viscosidad de 91 cps y una fuerza iónica de 0.15 M para imitar un ambiente acuoso. Se utilizó el método SHAKE para mantener todos los enlaces de los átomos de hidrógeno rígidos. La generación de la lista de vecinos se calculó cada 10 pasos con un corte de 14 Å, y las interacciones no covalentes se calcularon usando un corte de 12 Å para las interacciones de Coulomb y una función switch de suavizado en 10 Å para las interacciones de Van der Waals. A cada modelo se le aplicaron 2000 pasos de minimización estructural al vacío para eliminar choques

estéricos antes de la simulación de MD. Cada modelo de la región S2 se simuló durante 2 μs y las trayectorias fueron almacenadas cada 1 ps. La suma de las trayectorias de los 27 modelos genera un ensamble total de 54 μs.

Las simulaciones de dinámica molecular de las mutantes se realizaron en las mismas condiciones que las estructuras del ensamble inicial. Cada uno de los modelos se simuló durante 6 μs. Así mismo, las simulaciones de los modelos utilizados para generar las mutantes en las condiciones silvestres se extendieron a 6 μs. La suma de las trayectorias por variante genera un ensamble total de 24 μs. Un resumen de las simulaciones se muestra en la Tabla 5.

4.7 Validación de las simulaciones de dinámica molecular con CHARMM36 y solvente implícito

4.7.1 Simulación de $(AAQAA)_3$ helicoidal y extendido

Se generó el modelo de $(AAQAA)_3$ completamente extendido y helicoidal utilizando el programa Chimera[98]. Para cada modelo, el N-terminal y C-terminal fueron neutralizados por un grupo acetil (ACE) y por un grupo amida (CT2), respectivamente. Se utilizó CHARMM-GUI para generar los archivos de entrada para su simulación. Cada uno de los modelos se simuló de la misma manera que los modelos de la región S2 de Esg. Cada modelo de $(AAQAA)_3$ se simuló durante 16 μs (Tabla 5).

4.7.2 Caracterización del muestreo de hélices α_L

Se evaluaron los ángulos diedros phi (ϕ) y psi (φ) de los 54 μs del ensamble silvestre de S2. Se seleccionaron los ángulos ϕ y φ que estuvieran en el intervalo de un α_L. Un α_L es definida teniendo al menos tres residuos consecutivos con valores de ϕ y φ que caen en la región α_L ($30° < \phi < 100°$ and $7° < \varphi < 67°$)[99]. La probabilidad de α_L se calcula como la fracción del ensamble que contiene hélices α_L.

4.7.3 Monitoreo de la convergencia estructural

Para determinar la heterogeneidad estructural del ensamble inicial de los modelos de la región S2 de *Dme*-Esg, se calculó el RMSD de los átomos C_α entre todas las estructuras iniciales con charmm38b1. Posteriormente, para determinar si fue suficiente el tiempo de simulación, se realizó una matriz de RMSD en dos dimensiones (RMSD 2D) de los átomos C_α entre pares de trayectorias de cada uno de los modelos de la región S2 obtenidas de la

simulación de MD, y se identificó la distancia de RMSD más pequeña entre cada par de trayectorias. La matriz de RMSD 2D se generó con PTRAJ[100] (Process TRAJectory, en inglés), el cual es un programa perteneciente a las herramientas de análisis de Amber. Posteriormente se graficó la matriz de RMSD 2D de las trayectorias de los modelos que mostraron el valor de RMSD mayor y menor entre las distancias mínimas de RMSD.

4.8 Simulaciones de dinámica molecular del α-MoRF propuesto en la región S2 de *Dme*-Esg con CHARMM36m y solvente explícito

Se generó un modelo del α-MoRF (residuos 120 al 152 de la región S2 de *Dme*-Esg) completamente helicoidal utilizando el programa Chimera[98]. Se utilizó CHARMM-GUI para construir la variante silvestre y las mutantes E141P y E141L, las cuales se sintetizaron y se caracterizaron por dicroísmo circular en los laboratorios del Dr. Carlos Amero y la Dra. Lina Rivillas.

Cada modelo fue neutralizado en su extremo C-terminal por un grupo amida (CT2). Las cajas fueron generadas con CHARMM-GUI, las cuales fueron rectangulares y con una distancia de la proteína con respecto al borde de 12 Å. Todas las simulaciones de MD se desarrollaron usando el programa GROMACS[101] versión 2019.2 (URL http://manual.gromacs.org/documentation/2019/download.html), aplicando el campo de fuerza CHARMM36m[99], usando como solvente explícito el modelo de agua TIP3P y una concentración de sal de 0.15 NaCl. Las simulaciones se realizaron utilizando los parámetros por default sugeridos por CHARMM-GUI, en un ensamble NPT, a temperatura (300 K) y presión (1atm) constante utilizando un termostato de rescalamiento de velocidades (V-rescale) y el pistón Parrinello-Rahman, respectivamente. Se utilizó el método de LINCS para restringir los enlaces covalentes de átomos unidos a hidrógenos. Las interacciones electrostáticas de largo alcance fueron calculadas usando el método de las sumas de Ewald (PME) con un corte de 12 Å. Las interacciones no covalentes se calcularon usando un corte de 12 Å para las interacciones de Coulomb y una función switch de suavizado en 10 Å para las interacciones de Van der Waals. La lista de vecinos se recalculó cada 20 pasos con un radio de corte en 12 Å. Los archivos de entrada de las simulaciones fueron generados con CHARMM-GUI con los parámetros anteriormente mencionados y bajo el esquema de repartición de masas para átomos de hidrógeno (Hydrogen mass repartitioning, HMR, en inglés)[102,103]. Este método consiste en incrementar la masa de los átomos de hidrógeno hasta ~ 3 uma y disminuir lo suficiente la masa de los átomos pesados unidos a átomos de hidrógeno para mantener en el sistema una masa constante, de tal manera que esto permite incrementar el paso de integración de una simulación de MD hasta 4 fs y obtener resultados

razonables. El tiempo de integración para nuestras simulaciones fue de 4 fs y la frecuencia de guardado de 1 ps. A cada modelo se le aplicaron 10,000 pasos de minimización estructural y 500,000 pasos de equilibración con el solvente, esto con el objetivo de eliminar choques estéricos antes de la simulación de MD. Se generó para cada variante tres simulaciones independientes utilizando su estructura inicial; el tiempo que se simuló cada trayectoria y el tiempo acumulado para cada variante se muestra en la Tabla 4 y Tabla 5. Se generó un ensamble de cada variante considerando de cada trayectoria independiente la primera conformación que presentó 0 % de α-hélice al menos durante 500ps. De este ensamble, se realizó el análisis de fracción helicoidal, estructura secundaria por residuo y contactos sobre carbonos.

Tabla 4. Tiempo de simulación para cada una de las trayectorias independientes de la variante silvestre y las mutantes E141P y E141L.

Modelo	Trayectoria 1	Trayectoria 2	Trayectoria 3	Tiempo acumulado
silvestre	15 μs	9.800 μs	5 μs	29.8 μs
E141L	19.5 μs	14 μs	5 μs	38.5 μs
E141P	19.5 μs	5 μs	10.175 μs	34.675 μs

4.9 Validación de las simulaciones de dinámica molecular con CHARMM36m y solvente explícito

4.9.1 Simulación de $(AAQAA)_3$ helicoidal

Se generó el modelo de $(AAQAA)_3$ completamente helicoidal utilizando el programa Chimera[98]. El extremo C-terminal fue neutralizado por un grupo amida (CT2). Se utilizó CHARMM-GUI para generar los archivos de entrada para su simulación. El modelo de $(AAQAA)_3$ se simuló durante 16 *μ*s de la misma manera que los modelos de la región α-MoRF de *Dme*-Esg simulados con solvente explícito y CHARMM36m en GROMACS. Un resumen de las simulaciones realizadas con solvente explícito se muestra en la Tabla 5.

4.10 Cálculo de propiedades estructurales

Las trayectorias fueron analizadas utilizando CARMA[87] para calcular estructura secundaria, y charmm38b1 para realizar cálculos de desviación cuadrática media (RMSD), radio de giro (Rg), ángulos diedros, contactos y puentes de hidrógeno. Las trayectorias obtenidas con Gromacs y solvente explícito, se cambiaron de formato xtc a dcd con la librería de mdconvert del programa

MDTraj[104] para ser analizadas con charmm38b1. Los resultados fueron graficados con Matplotlib 2.2.3[79,80].

5.- Resultados y Discusión

5.1 Análisis bioinformático de los factores de transcripción de la familia Snail

5.1.1 Análisis de divergencia en longitud y secuencia

Se ha reportado que el N-terminal de los factores de transcripción de la familia Snail es divergente. Para conocer la divergencia que presenta el factor de transcripción Esg expresado en *Dme* y sus ortólogos, se realizó un blast a nivel de proteína y se obtuvieron secuencias de factores de transcripción de la familia Snail (snai1 y snai2) las cuales se expresan en diferentes especies y a su vez pertenecen a diferentes filos (Tabla 2 y Tabla 6). Así mismo, un mapa filogenético (Figura 7) fue construido con la herramienta de Taxonomy browser de NCBI, en el cual se pueden ver representados los diferentes filos a los que pertenecen Esg de *Dme* y sus ortólogos. *Amphimedon queenslandica* presenta una proteína no caracterizada y es un caso interesante de analizar ya que pertenece al filo de *Porifera*[105], el cual incluye a las esponjas y representa a los primeros ancestros de los metazoarios, con las características principales de ser organismos con una fase unicelular, tener un cuerpo no segmentado y no presentar tejidos.

Tabla 5. Resumen del total de simulaciones que se realizaron en este proyecto

Región S2 de Esg				
Modelo	**Solvente**	**Campo de fuerza**	**Tiempo simulado**	**Tiempo acumulado**
*27 Modelos diferentes (versión silvestre)	Implícito	CHARMM36	2μs (por modelo)	54μs
*HHP, PI, SP6 y SP9 (versión silvestre)			6μs (por modelo)	24μs
*HHP, PI, SP6 y SP9 (mutante E141P)				24μs
*HHP, PI, SP6 y SP9 (mutante E141L)				24μs
*HHP, PI, SP6 y SP9 (mutante E141V)				24μs
*HHP, PI, SP6 y SP9 (mutante E141A_T145A)				24μs
*HHP, PI, SP6 y SP9 (mutante Q149P)				24μs
*HHP, PI, SP6 y SP9 (mutante R144A)				24μs
*HHP, PI, SP6 y SP9 (mutante T145V)				24μs
Helicoidal-CT2 (versión silvestre)	Explícito	CHARMM36m	Trayectoria 1: 15 μs	29.8 μs
			Trayectoria 2: 9.8 μs	
			Trayectoria 3: 15 μs	
Helicoidal-CT2 (mutante E141L)			Trayectoria 1: 19.5 μs	38.5 μs
			Trayectoria 2: 14 μs	
			Trayectoria 3: 5 μs	
Helicoidal-CT2 (mutante E141P)			Trayectoria 1: 19.5 μs	34.675 μs
			Trayectoria 2: 5 μs	
			Trayectoria 3: 10.175 μs	
Molécula control: $(AAQAA)_3$				
Modelo	**Solvente**	**Campo de fuerza**	**Tiempo simulado**	**Tiempo acumulado**
Helicoidal ACE-$(AAQAA)_3$-CT2	Implícito	CHARMM36	16 μs	32 μs
Extendido ACE-$(AAQAA)_3$-CT2			16 μs	
Helicoidal $(AAQAA)_3$-CT2	Explícito	CHARMM36m	16 μs	16 μs

Nota:
* Modelos obtenidos del ensamble inicial de la región S2 de Esg (Figura 17)

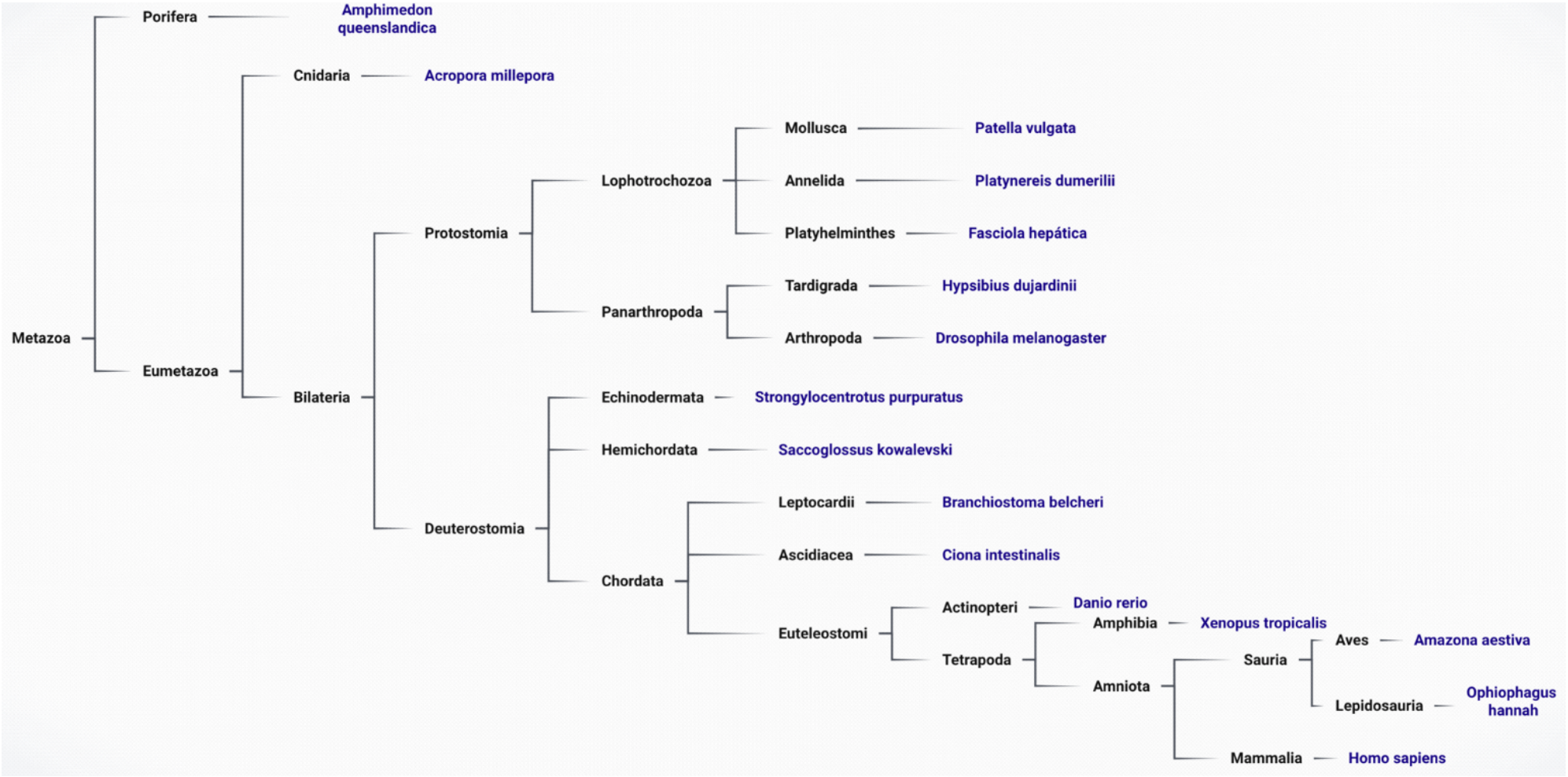


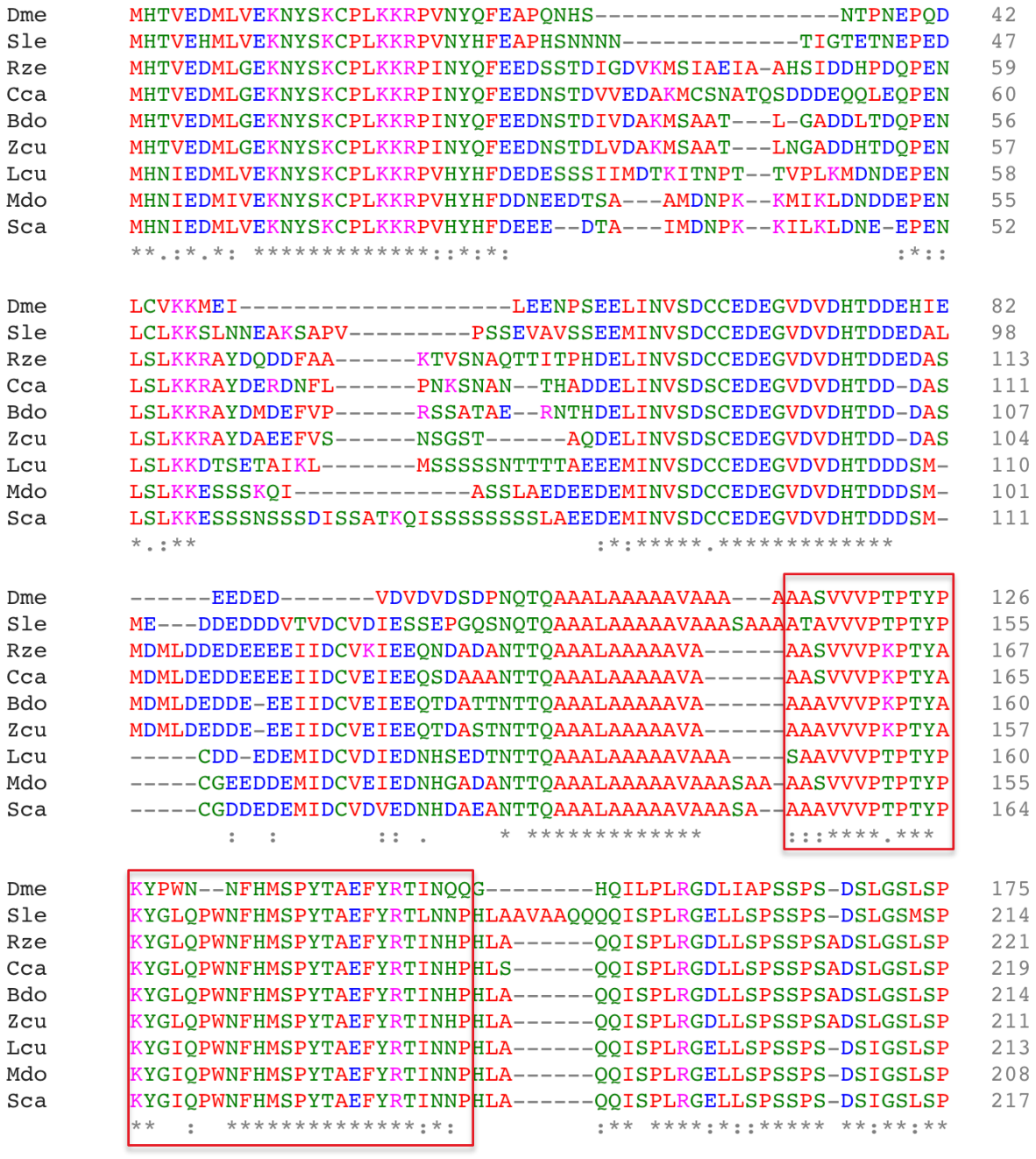
Figura 7. Mapa filogenético del factor de transcripción de Esg de *Dme* y sus ortólogos.

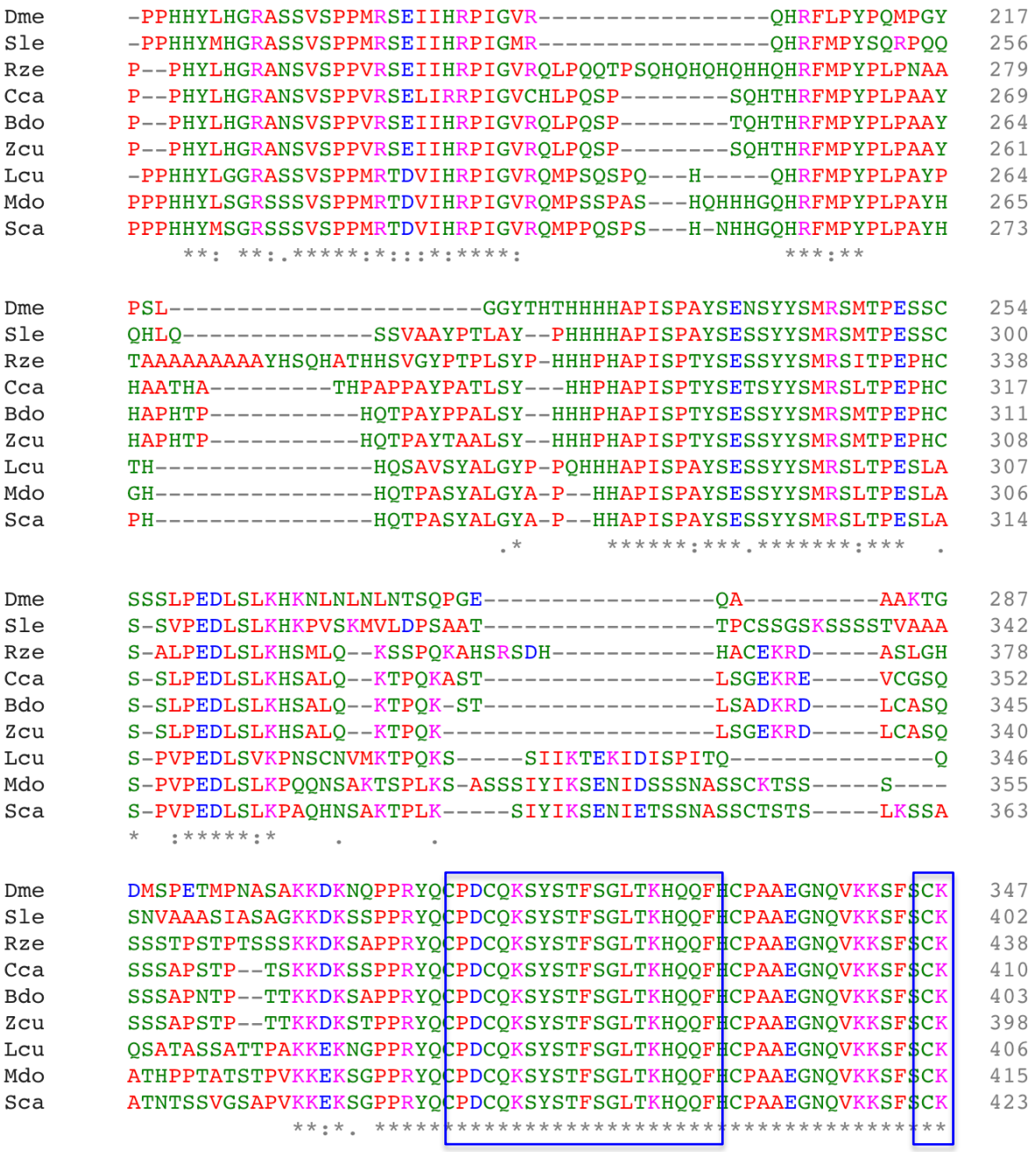
En el mapa filogenético se puede ver representado el filo *Eumetazoa* que comprende los filos *Cnidaria* y *Bilateria*. El filo *Cnidaria* contiene organismos acuáticos que presentan una simetría radial; a este grupo pertenece *Acropora millepora*. Por su parte, el filo *Bilateria* contiene organismos con un cuerpo segmentado en dos ejes simétricos, presentan tejidos y comprende organismos que desarrollan primero la boca (filo *Protostomia*) o el ano (filo *Deuterostomia*) durante el desarrollo embrionario[106]. El filo *Protostomia* incluye a los invertebrados como moluscos (*Patella vulgata*), anélidos (*Platynereis dumerilii*), platelmitos (*Fasciola hepática*), tardígrados (*Hypsibius dujardinii*) y artrópodos (*Drosophila melanogaster*). El filo *Deuterostomia* comprende a organismos con una organización más compleja, los cuales pueden ser vertebrados o invertebrados. Por ejemplo, el filo *Echinodermata*[107] contiene organismos invertebrados que presentan una simetría pentaradial y un cuerpo espinoso, presentan tejido conectivo y un sistema vascular de agua que les permite tener movilidad; a este grupo pertenece *Strongylocentrotus purpuratus*. A su vez, los organismos que pertenecen al filo *Hemichordata*[108] presentan hendiduras branquiales y carecen de un sistema nervioso desarrollado, ya que presentan un cordón nervioso dorsal hueco; a este grupo pertenece *Saccoglossus kowalevski*. Finalmente, los organismos que pertenecen al filo *Chordata*[109] presentan notocorda, la cual es una estructura flexible que proporciona soporte esquelético a lo largo del cuerpo y en organismos vertebrados es reemplazada por la columna vertebral; también presentan sistema nervioso. En este grupo encontramos a leptocardios (*Branchiostoma belcheri),* ascidias (*Ciona intestinalis*), actinopterigios (*Danio rerio*), anfibios (*Xenopus tropicalis*), aves (*Amazona aestiva*), lepidosaurios (*Ophiophagus hannah*) y mamíferos (*Homo sapiens*). Es importante mencionar que las proteínas de *Amphimedon queenslandica* y *Branchiostoma belcheri* no está claro si pertenecen a la familia Snail; sin embargo, presentan ZNFs que les permite alinear con el extremo C-terminal de *Dme*-Esg, y por esta razón fueron consideradas.

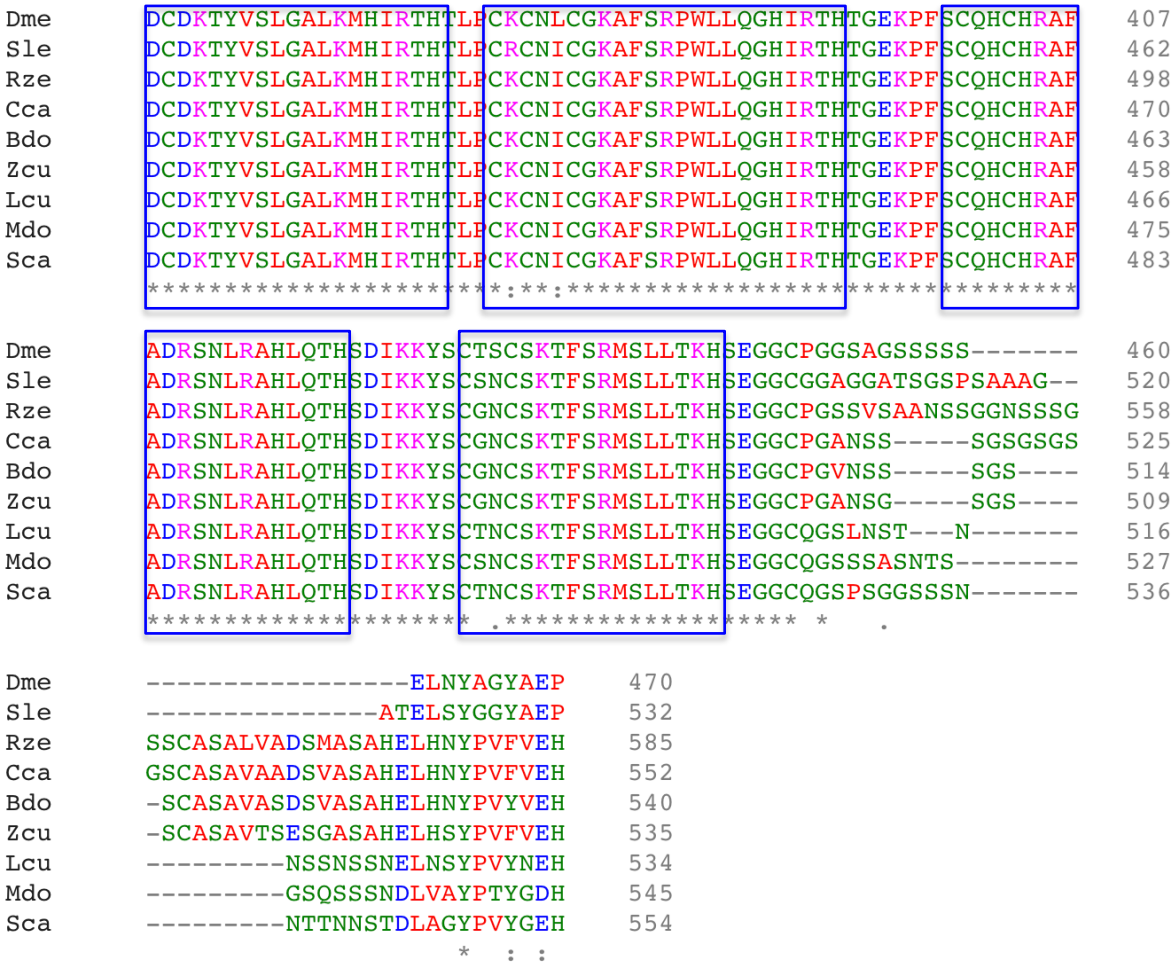
En la Tabla 6 se puede observar que las secuencias de Esg de *Dme* y sus ortólogos en general presentan diferente longitud, lo cual se observa en las variaciones de longitud que presentan sus extremos N-terminal y C-terminal. Todas las secuencias de los ortólogos muestran diferente porcentaje de identidad en secuencia respecto a *Dme*-Esg, la cual oscila entre 41.81 % y 79.56 %, centrada en los dedos de zinc. De igual forma, se realizó un blast utilizando el N-terminal de *Dme*-Esg como anzuelo para identificar proteínas homólogas; las secuencias obtenidas se pueden observar en la Tabla 7, las cuales comparten un porcentaje de identidad que oscila entre 49.37 % y 64.04 %. Es importante mencionar que ninguna de las secuencias (Tabla 6 y Tabla 7) cuenta con estructuras reportadas en el Protein Data Bank (PDB).

Utilizando Clustal Omega, se realizaron los alineamientos de secuencia. En general, los resultados mostraron que Esg de *Dme* y sus ortólogos comparten un extremo C-terminal bastante similar, al ser la región donde se expresan los ZNFs y cuyo porcentaje de identidad obtenido y mostrado en la Tabla 6 corresponde a esta región. Sin embargo, las secuencias de Esg de *Dme* y sus ortólogos fueron incapaces de alinearse en su extremo N-terminal, lo que sugiere que son divergentes a nivel de secuencia. Por su parte, el alineamiento de Esg de *Dme* y sus homólogos en moscas (Figura 8) muestra que comparten un extremo N-terminal y C-terminal bastante similares a nivel de secuencia.

Figura 8. Alineamiento de secuencia del factor de transcripción Esg de *Dme* y proteínas homólogas, obtenido con Clustal Omega. En el cuadro rojo se representa la región S2 de *Dme*-Esg (116-148 aa) presente en el N-terminal, la cual está conservada en proteínas homólogas; mientras que en el cuadro azul se representa la región de ZNFs presente en el C-terminal.




Dme -PPHHYLHGRASSVSPPMRSEIIHRPIGVR----------QHRFLPYPQMPGY 217
Sle -PPHHYMHGRASSVSPPMRSEIIHRPIGMR----------QHRFMPYSQRPQQ 256
Rze P--PHYLHGRANSVSPPVRSEIIHRPIGVRQLPQQTPSQHQHQHQHHQHRFMPYPLPNAA 279
Cca P--PHYLHGRANSVSPPVRSELIRRPIGVCHLPQSP--------SQHTHRFMPYPLPAAY 269
Bdo P--PHYLHGRANSVSPPVRSEIIHRPIGVRQLPQSP--------TQHTHRFMPYPLPAAY 264
Zcu P--PHYLHGRANSVSPPVRSEIIHRPIGVRQLPQSP--------SQHTHRFMPYPLPAAY 261
Lcu -PPHHYLGGRASSVSPPMRTDVIHRPIGVRQMPSQSPQ---H-----QHRFMPYPLPAYP 264
Mdo PPPHHYLSGRSSSVSPPMRTDVIHRPIGVRQMPSSPAS---HQHHHGQHRFMPYPLPAYH 265
Sca PPPHHYMSGRSSSVSPPMRTDVIHRPIGVRQMPPQSPS---H-NHHGQHRFMPYPLPAYH 273
PSL----------GGYTHTHHHHAPISPAYSENSYYSMRSMTPESSC 254
QHLQ----------SSVAAYPTLAY--PHHHHAPISPAYSESSYYSMRSMTPESSC 300
TAAAAAAAAAYHSQHATHHSVGYPTPLSYP-HHHPHAPISPTYSESSYYSMRSITPEPHC 338
HAATHA---------THPAPPAYPATLSY---HHPHAPISPTYSETSYYSMRSLTPEPHC 317
HAPHTP----------HQTPAYPPALSY--HHHPHAPISPTYSESSYYSMRSMTPEPHC 311
HAPHTP----------HQTPAYTAALSY--HHHPHAPISPTYSESSYYSMRSMTPEPHC 308
TH----------HQSAVSYALGYP-PQHHHAPISPAYSESSYYSMRSLTPESLA 307
GH----------HQTPASYALGYA-P--HHAPISPAYSESSYYSMRSLTPESLA 306
PH----------HQTPASYALGYA-P--HHAPISPAYSESSYYSMRSLTPESLA 314
SSSLPEDLSLKHKNLNLNLNTSQPGE----------QA----------AAKTG 287
S-SVPEDLSLKHKPVSKMVLDPSAAT----------TPCSSGSKSSSSTVAAA 342
S-ALPEDLSLKHSMLQ--KSSPQKAHSRSDH----------HACEKRD-----ASLGH 378
S-SLPEDLSLKHSALQ--KTPQKAST----------LSGEKRE-----VCGSQ 352
S-SLPEDLSLKHSALQ--KTPQK-ST----------LSADKRD-----LCASQ 345
S-SLPEDLSLKHSALQ--KTPQK----------LSGEKRD-----LCASQ 340
S-PVPEDLSVKPNSCNVMKTPQKS-----SIIKTEKIDISPITQ----------Q 346
S-PVPEDLSLKPQQNSAKTSPLKS-ASSSIYIKSENIDSSSNASSCKTSS-----S---- 355
S-PVPEDLSLKPAQHNSAKTPLK-----SIYIKSENIETSSNASSCTSTS-----LKSSA 363
DMSPETMPNASAKKDKNQPPRYQCPDCQKSYSTFSGLTKHQQFHCPAAEGNQVKKSFSCK 347
SNVAAASIASAGKKDKSSPPRYQCPDCQKSYSTFSGLTKHQQFHCPAAEGNQVKKSFSCK 402
SSSTPSTPTSSSKKDKSAPPRYQCPDCQKSYSTFSGLTKHQQFHCPAAEGNQVKKSFSCK 438
SSSAPSTP--TSKKDKSSPPRYQCPDCQKSYSTFSGLTKHQQFHCPAAEGNQVKKSFSCK 410
SSSAPNTP--TTKKDKSAPPRYQCPDCQKSYSTFSGLTKHQQFHCPAAEGNQVKKSFSCK 403
SSSAPSTP--TTKKDKSTPPRYQCPDCQKSYSTFSGLTKHQQFHCPAAEGNQVKKSFSCK 398
QSATASSATTPAKKEKNGPPRYQCPDCQKSYSTFSGLTKHQQFHCPAAEGNQVKKSFSCK 406
ATHPPTATSTPVKKEKSGPPRYQCPDCQKSYSTFSGLTKHQQFHCPAAEGNQVKKSFSCK 415
ATNTSSVGSAPVKKEKSGPPRYQCPDCQKSYSTFSGLTKHQQFHCPAAEGNQVKKSFSCK 423

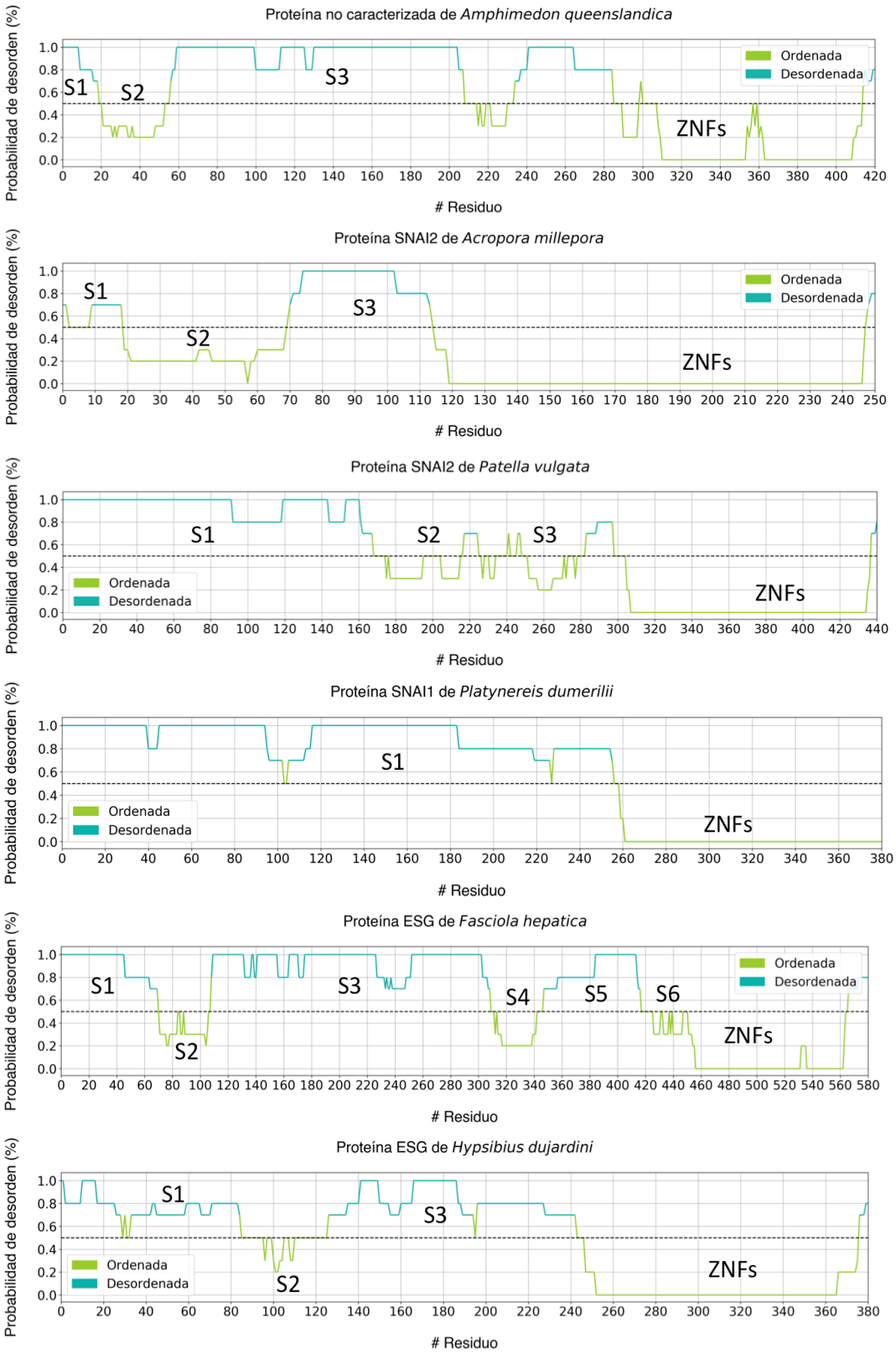

Dme DCDKTYVSLGALKMHIRTHTLPCKCNLCGKAFSRPWLLQGHIRTHTGEKPFSCQHCHRAF 407
Sle DCDKTYVSLGALKMHIRTHTLPCRCNICGKAFSRPWLLQGHIRTHTGEKPFSCQHCHRAF 462
Rze DCDKTYVSLGALKMHIRTHTLPCKCNICGKAFSRPWLLQGHIRTHTGEKPFSCQHCHRAF 498
Cca DCDKTYVSLGALKMHIRTHTLPCKCNICGKAFSRPWLLQGHIRTHTGEKPFSCQHCHRAF 470
Bdo DCDKTYVSLGALKMHIRTHTLPCKCNICGKAFSRPWLLQGHIRTHTGEKPFSCQHCHRAF 463
Zcu DCDKTYVSLGALKMHIRTHTLPCKCNICGKAFSRPWLLQGHIRTHTGEKPFSCQHCHRAF 458
Lcu DCDKTYVSLGALKMHIRTHTLPCKCNICGKAFSRPWLLQGHIRTHTGEKPFSCQHCHRAF 466
Mdo DCDKTYVSLGALKMHIRTHTLPCKCNICGKAFSRPWLLQGHIRTHTGEKPFSCQHCHRAF 475
Sca DCDKTYVSLGALKMHIRTHTLPCKCNICGKAFSRPWLLQGHIRTHTGEKPFSCQHCHRAF 483
ADRSNLRAHLQTHSDIKKYSCTSCSKTFSRMSLLTKHSEGGCPGGSAGSSSSS------- 460
ADRSNLRAHLQTHSDIKKYSCSNCSKTFSRMSLLTKHSEGGCGGAGGATSGSPSAAAG-- 520
ADRSNLRAHLQTHSDIKKYSCGNCSKTFSRMSLLTKHSEGGCPGSSVSAANSSGGNSSSG 558
ADRSNLRAHLQTHSDIKKYSCGNCSKTFSRMSLLTKHSEGGCPGANSS-----SGSGSGS 525
ADRSNLRAHLQTHSDIKKYSCGNCSKTFSRMSLLTKHSEGGCPGVNSS-----SGS---- 514
ADRSNLRAHLQTHSDIKKYSCGNCSKTFSRMSLLTKHSEGGCPGANSG-----SGS---- 509
ADRSNLRAHLQTHSDIKKYSCTNCSKTFSRMSLLTKHSEGGCQGSLNST---N------- 516
ADRSNLRAHLQTHSDIKKYSCSNCSKTFSRMSLLTKHSEGGCQGSSSASNTS-------- 527
ADRSNLRAHLQTHSDIKKYSCTNCSKTFSRMSLLTKHSEGGCQGSPSGGSSSN------- 536
----------ELNYAGYAEP 470
----------ATELSYGGYAEP 532
SSCASALVADSMASAHELHNYPVFVEH 585
GSCASAVAADSVASAHELHNYPVFVEH 552
-SCASAVASDSVASAHELHNYPVYVEH 540
-SCASAVTSESGASAHELHSYPVFVEH 535
---------NSSNSSNELNSYPVYNEH 534
---------GSQSSSNDLVAYPTYGDH 545
---------NTTNNSTDLAGYPVYGEH 554

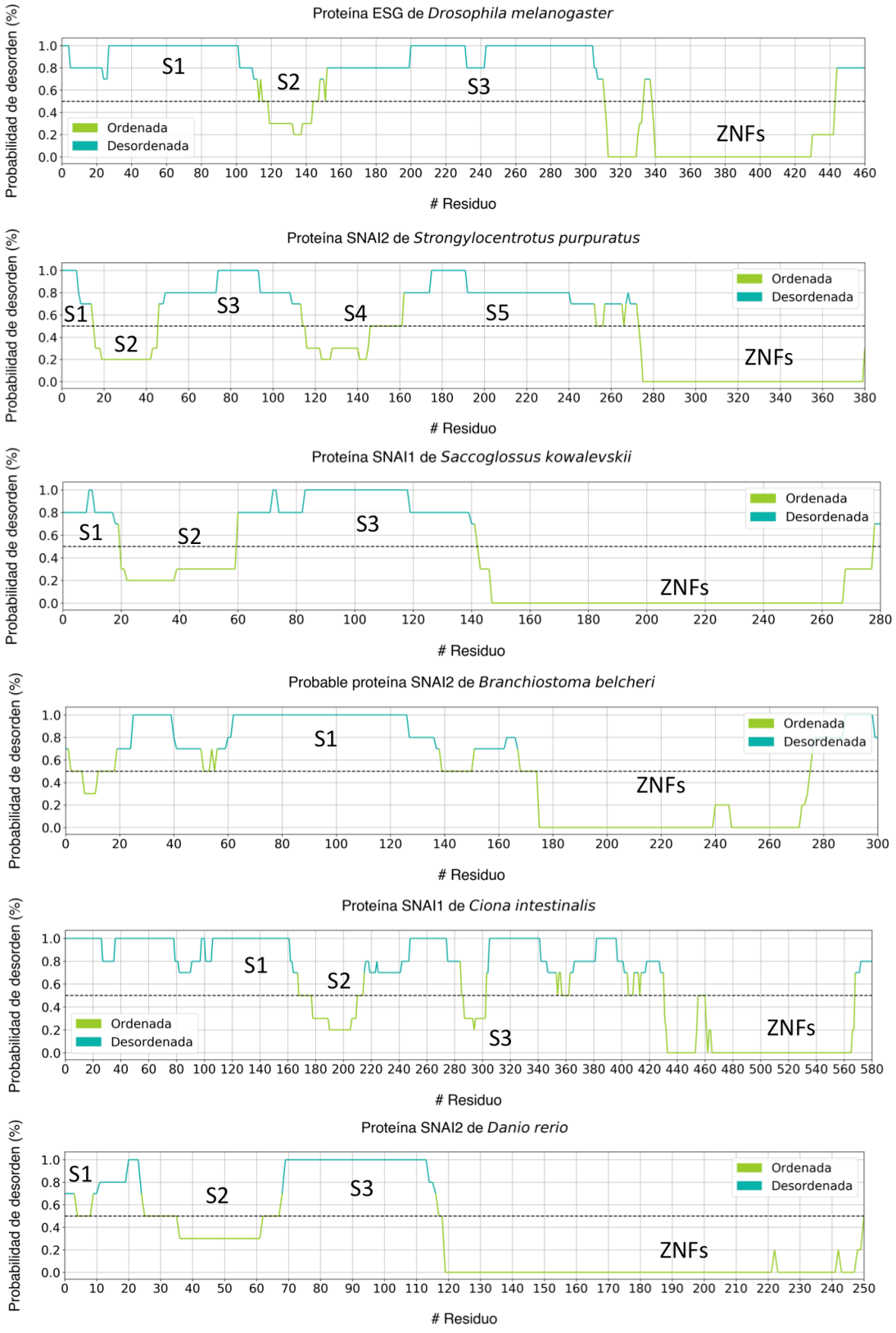
5.1.2 Análisis de desorden: Existencia de regiones con tendencia a ordenarse

Muchos de los factores de transcripción son proteínas que presentan regiones desordenadas[5]. Se ha reportado que las regiones desordenadas pueden presentar conservación a nivel de secuencia, a nivel de desorden y/o funciones conservadas independientemente del resto de la cadena polipeptídica[85]. Para conocer el grado de desorden que presentan el factor de transcripción Esg de *Dme* y sus ortólogos, se determinó la probabilidad de orden y desorden a nivel de secuencia, utilizando diferentes predictores de desorden: Metadisorder, MDFP2, AUCpreD y SPOT-disorder. Los resultados obtenidos de los predictores se promediaron y se obtuvo una gráfica del perfil de desorden por proteína.

Las gráficas de desorden (Figura 9) muestran regiones ordenadas y desordenadas de diferente longitud. El extremo C-terminal de Esg de *Dme* y sus ortólogos es la región donde se encuentran los ZNFs, la cual está bien caracterizada estructural y funcionalmente, y se muestra como una región ordenada, a diferencia del extremo N-terminal, el cual presenta regiones ordenadas y desordenadas. En la Tabla 8 se muestran las longitudes de las regiones propensas a ordenarse presentes en el N-terminal de Esg de *Dme* y sus ortólogos. Aquellas regiones predichas como ordenadas que estuvieran presentes en el N-terminal y tuvieran una longitud mínima de 30 aminoácidos, se consideraron como regiones independientes de las regiones desordenadas anexas a ella. Estás regiones se identificaron en color violeta en la Tabla 8.

Figura 9. Perfil de Desorden del factor de transcripción de Esg de *Dme* y sus ortólogos.




Proteína ESG de *Drosophila melanogaster*
Probabilidad de desorden (%)
S1
S2
S3
ZNFs
Ordenada
Desordenada
Residuo
Proteína SNAI2 de *Strongylocentrotus purpuratus*
Probabilidad de desorden (%)
S1
S2
S3
S4
S5
ZNFs
Ordenada
Desordenada
Residuo
Proteína SNAI1 de *Saccoglossus kowalevskii*
Probabilidad de desorden (%)
S1
S2
S3
ZNFs
Ordenada
Desordenada
Residuo
Probable proteína SNAI2 de *Branchiostoma belcheri*
Probabilidad de desorden (%)
S1
ZNFs
Ordenada
Desordenada
Residuo
Proteína SNAI1 de *Ciona intestinalis*
Probabilidad de desorden (%)
S1
S2
S3
ZNFs
Ordenada
Desordenada
Residuo
Proteína SNAI2 de *Danio rerio*
Probabilidad de desorden (%)
S1
S2
S3
ZNFs
Ordenada
Desordenada
Residuo

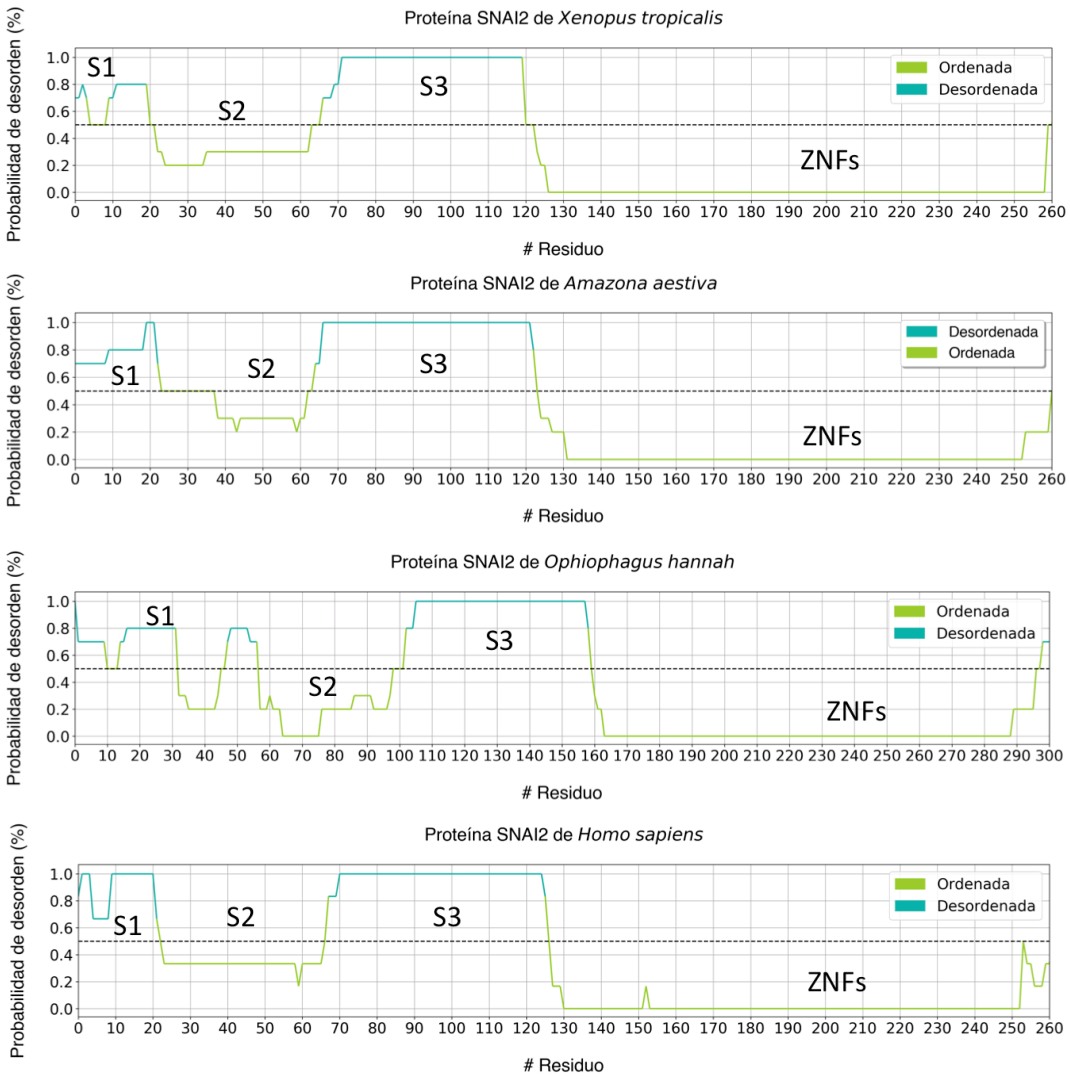

Proteína SNAI2 de *Xenopus tropicalis*
Probabilidad de desorden (%)
S1
S2
S3
ZNFs
Ordenada
Desordenada
Residuo
Proteína SNAI2 de *Amazona aestiva*
Probabilidad de desorden (%)
S1
S2
S3
ZNFs
Desordenada
Ordenada
Residuo
Proteína SNAI2 de *Ophiophagus hannah*
Probabilidad de desorden (%)
S1
S2
S3
ZNFs
Ordenada
Desordenada
Residuo
Proteína SNAI2 de *Homo sapiens*
Probabilidad de desorden (%)
S1
S2
S3
ZNFs
Ordenada
Desordenada
Residuo

Tabla 6. Factor de transcripción Esg de *Dme* y proteínas ortólogas seleccionadas a través de un blast a nivel de proteína considerando la secuencia de ZNFs de *Dme*

Organismo	Abreviación	Proteína	Filo	Longitud (aa)	Identificador	N-terminal	C-terminal
Amphimedon queenslandica	*Aqu*	**	*Porifera*	431	XP_019864429.1	1-285	286-431
Acropora millepora	*Ami*	snai2	*Cnidaria*	257	ACO55053	1-115	116-257
Patella vulgata	*Pvu*	snai2	*Mollusca*	444	AAL12167.1	1-302	303-444
Platynereis dumerilii	*Pdu*	snai1	*Annelida*	400	CAX51846.1	1-258	259-400
Fasciola hepatica	*Fhe*	esg	*Plathelmintos*	588	THD25286.1	1-458	459-588
Hypsibius dujardini	*Hdu*	esg	*Tardigrada*	398	OQV12667.1	1-246	247-398
Drosophila melanogaster	*Dme*	esg	*Arthropoda*	470	P25932	1-308	309-470
Strongylocentrotus purpuratus	*Spu*	snai2	*Echinodermata*	385	XP_785413	1-244	245-385
Saccoglossus kowalevskii	*Sko*	snai1	*Hemichordata*	287	NP_001158460.1	1-144	145-287
Branchiostoma belcheri	*Bbe*	*	*Chordata*	310	XP_019629674.1	1-168	169-310
Ciona intestinalis	*Cin*	snai1	*Chordata*	584	AAB61226.1	1-435	436-584
Danio rerio	*Dre*	snai2	*Chordata*	257	NP_001008581	1-116	117-257
Xenopus tropicalis	*Xtr*	snai2	*Chordata*	266	NP_989424	1-125	126-266
Amazona aestiva	*Aae*	snai2	*Chordata*	268	KQK73718	1-127	128-268
Ophiophagus hannah	*Oha*	snai2	*Chordata*	304	ETE62258	1-163	164-304
Homo sapiens	*Hsa*	snai2	*Chordata*	268	NP_003059	1-127	128-268

Nota:
** Proteína no caracterizada
*Probable proteína snai2

Tabla 7. Factor de transcripción Esg de *Dme* y proteínas homólogas seleccionadas a través de un blast considerando la secuencia completa de *Dme*.

Organismo		Proteína	Longitud (aa)	Identificador	Identidad (%)	Cobertura (%)
Drosopholila melanogaster	*Dme*	esg	470	NP_476600.1	-	-
Scaptodrosophila lebanonensis	*Sle*		532	XP_030383794.1	64.04	86
Bactrocera dorsalis	*Bdo*		540	XP_011210071.1	53.99	100
Zeugodacus cucurbitae	*Zcu*		535	XP_011189395.1	54.18	100
Lucilia cuprina	*Lcu*		534	XP_023292153.1	53.23	100
Ceratis capitata	*Cca*		552	XP_004531537.1	51.19	100
Rhagoletis zephyria	*Rze*	*	585	XP_017470371.1	49.39	100
Stomoxys calcitrans	*Sca*	*	554	XP_013108397.1	49.37	100
Musca domestica	*Mdo*	*	545	XP_005190299.1	49.74	100

*Probable proteína snai2

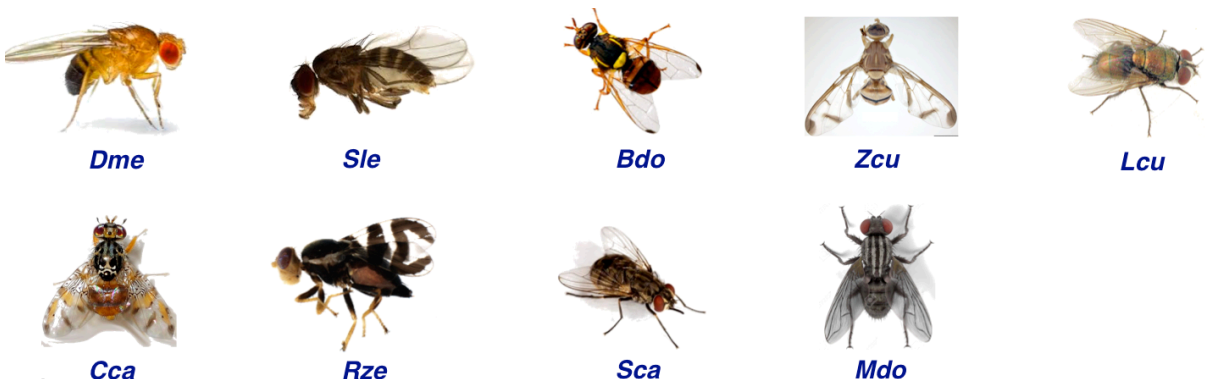


Figura 10. Especies de moscas utilizadas en el estudio bioinformático del factor de transcripción Esg de *Dme* y proteínas homólogas.

Tabla 8. Longitud de las regiones ordenadas presentes en el N-terminal de Esg de *Dme* y sus ortólogos. En color violeta se pueden identificar las regiones ordenadas que presentan una longitud mínima de 30 aminoácidos, así como el tipo de MoRF que podrían ser.

Organismo	Proteína	N-terminal	Regiones ordenadas (aa)	Probable tipo de MoRF
Aqu	**	1-285	20-56	β-MoRF
			209-234	-
Ami	snai2	1-115	1-9	-
			20-70	α-MoRF
Pvu	snai2	1-302	169-217	β-MoRF
			226-283	α-MoRF/β-MoRF
Pdu	snai1	1-258	-	-
Fhe	esg	1-458	71-108	α-MoRF
			310-347	α-MoRF
			418-458	β-MoRF
Hdu	esg	1-246	30-33	-
			86-126	β-MoRF
Dme	esg	1-308	116-148	α-MoRF
Spu	snai2	1-244	16-46	α-MoRF
			115-161	α-MoRF
Sko	snai1	1-144	21-60	α-MoRF
Bbe	*	1-168	3-19	-
			52-56	-
			140-151	-
Cin	snai1	1-435	169-215	α-MoRF
			286-303	-
			358-363	-
Dre	snai2	1-116	5-9	-
			26-68	α-MoRF
Xtr	snai2	1-125	5-9	-
			21-66	α-MoRF
Aae	snai2	1-127	24-64	α-MoRF
Oha	snai2	1-163	11-14	-
			33-47	-
			58-102	α-MoRF
Hsa	snai2	1-127	23-67	α-MoRF

Nota: ** Proteína no caracterizada, *Probable proteína snai2

De esta manera, se pudo identificar en la mayoría de las especies un patrón de desorden –orden- desorden presente en el N-terminal, las cuales hemos identificado como región S1 – región S2 – región S3, respectivamente (Figura 9). Sin embargo, hay especies que no presentan dicho patrón como *Fhe*, *Pvu* y *Spu* quienes presentan más de una región propensa a ordenarse en su extremo N-terminal. Así mismo, *Bbe* presenta islas pequeñas de regiones propensas a ordenarse, las cuales tienen una longitud entre 4 y 16 aminoácidos, mientras que *Pmu* presenta su N-terminal completamente desordenado. Los resultados sugieren que el extremo C-terminal de Esg de *Dme* y sus ortólogos es una región que se conserva a nivel de secuencia y estructura, a diferencia de su N-terminal, el cual podría conservarse únicamente a nivel de desorden.

5.1.3 Análisis de composición y tipo de IDR de regiones con tendencia a ordenarse

El factor de transcripción Esg de *Dme* y la mayoría de sus ortólogos mostraron un perfil de desorden similar, en el cual se identificaron regiones propensas a ordenarse presentes en el N-terminal. Se ha reportado que las IDPs pueden contener IDRs las cuales podrían presentar propiedades funcionales, químicas y estructurales independientes del resto de la proteína y podrían considerarse como dominios presentes en una región desordenada. Se analizó cada una de las secuencias (Tabla 6) con el servidor IDDomainSpotter[85], el cual considera la composición de secuencia para identificar IDRs que no adoptan estructura estable con un núcleo hidrofóbico. Este tipo de análisis es otra manera de identificar regiones conservadas en el N-terminal de Esg, así como caracterizar las regiones propensas a ordenarse (Tabla 8). En la Figura 11 se pueden observar los resultados obtenidos con IDDomainSpotter para *Dme*-Esg, que muestran que la región de ZNFs presente en el C-terminal de Esg es una región conservada entre *Dme* y sus ortólogos. La región de ZNFs está enriquecida de aminoácidos hidrofóbicos (+LVI) los cuales no son muy frecuentes en IDPs e IDRs a diferencia de las proteínas ordenadas. También mostró estar enriquecida de aminoácidos +RK y +PRK-DE, lo cual indica que es una región de unión a DNA por la presencia de los ZNFs, y a su vez, carece de aminoácidos +PST-RK, los cuales promueven desorden (Tabla 9). Estos resultados sugieren que la región de ZNFs de Esg de *Dme* y sus ortólogos es una zona estructurada. Por otra parte, se puede observar que el N-terminal de Esg de *Dme* y sus ortólogos es rico en aminoácidos +PST-RK, lo cual lo define como una zona flexible y desordenada, y

además carece de aminoácidos +RK y +PRK-DE, lo que indica que es una zona que no interactúa con ácidos nucleicos. A pesar de que se define el N-terminal como una región desordenada, hay regiones que son ricas en residuos hidrofóbicos (+LVI) y aromáticos (+FYG), principalmente en las regiones propensas a ordenarse (Tabla 8, Tabla 10) de los organismos que la presentan. El conjunto de los resultados sugiere que las regiones propensas a ordenarse son zonas con probables MoRFs (Tabla 8) que participen en funciones independientes a la región de ZNFs. Además, al estar enriquecidas con aminoácidos hidrofóbicos podrían adoptar estructuras compactas y la presencia de aminoácidos +PST-RK podría indicar que son potenciales sitios de fosforilación.

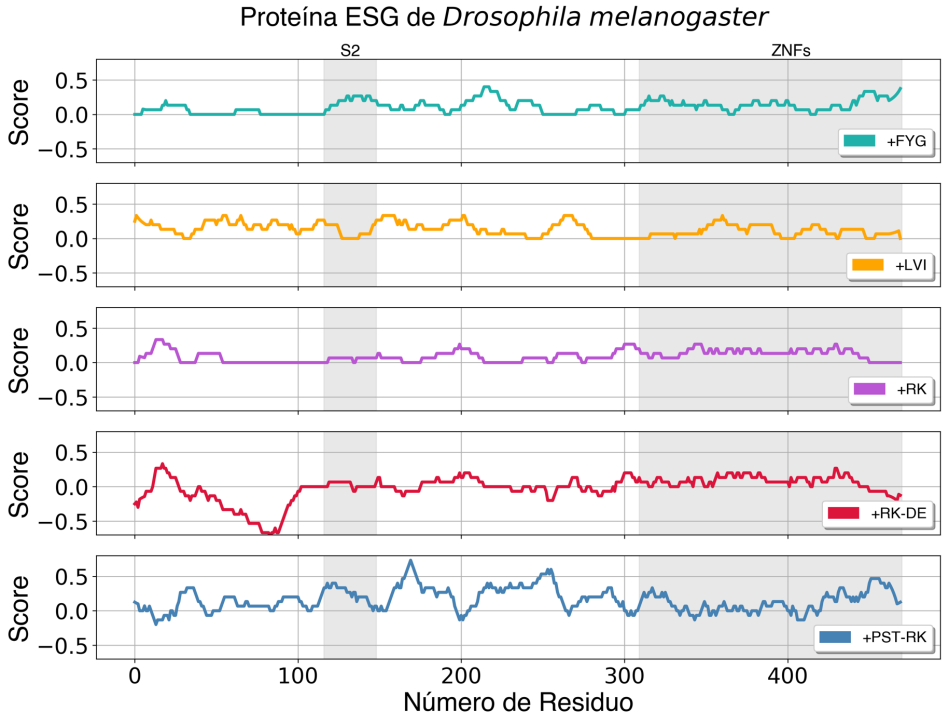


Figura 11. Diagrama obtenido por IDDomainSpotter del análisis de *Dme*-Esg con base en su composición. La región sombreada de color gris indica la región ZNFs, la cual corresponde al extremo C-terminal, así como la región S2 predicha como ordenada presente en el N-terminal. El diagrama muestra los scores para los residuos Phe+Tyr+Gly (+FYG), Leu+Val+Ile (+LVI), Arg+Lys-Asp-Glu (+RK-DE) y Pro+Ser+Thr-Arg-Lys (+PST-RK) calculados sobre una ventana de 15 residuos.

Tabla 9. Tablas de los valores obtenidos por IDDomainSpotter de la región ZNFs presente en el C-terminal de Esg de *Dme* y sus ortólogos. En color verde se indica la composición de aminoácidos que se presenta con mayor abundancia utilizando un corte de 0.1.

Aminoácidos	Dme (309-470 aa)		Ami (116-257 aa)		Dre (117-257 aa)		Hsa (128-268 aa)		Hdu (247-398 aa)		Cin (436-584 aa)	
	prom	σ	prom	σ	prom	σ	prom	σ	prom	σ	prom	σ
+FYG	0.15	0.08	0.13	0.07	0.12	0.07	0.12	0.06	0.12	0.05	0.09	0.06
+LVI	0.11	0.08	0.13	0.09	0.13	0.07	0.13	0.08	0.14	0.08	0.14	0.07
+RK	0.13	0.07	0.17	0.06	0.17	0.06	0.17	0.06	0.16	0.07	0.17	0.05
+RK-DE	0.07	0.09	0.12	0.08	0.10	0.09	0.10	0.09	0.12	0.07	0.09	0.07
+PST-RK	0.11	0.15	0.03	0.11	0.03	0.09	0.02	0.09	0.05	0.15	0.05	0.08

Aminoácidos	Oha (164-304)		Xtr (126-266 aa)		Aqu (268-431 aa)		Aes (128-268 aa)		Spu (245-385 aa)	
	prom	σ	prom	σ	prom	σ	prom	σ	prom	σ
+FYG	0.12	0.06	0.12	0.06	0.13	0.07	0.12	0.06	0.11	0.07
+LVI	0.13	0.08	0.13	0.08	0.11	0.06	0.13	0.08	0.13	0.07
+RK	0.17	0.06	0.17	0.06	0.17	0.07	0.17	0.06	0.18	0.06
+RK-DE	0.10	0.09	0.10	0.09	0.12	0.08	0.10	0.09	0.13	0.07
+PST-RK	0.02	0.09	0.03	0.10	0.09	0.13	0.02	0.09	0.03	0.13

Aminoácidos	Pvu (303-444 aa)		Fhe (459-588 aa)		Bbe (169-310 aa)		Sko (145-287 aa)		Pdu (259-400 aa)	
	prom	σ	prom	σ	prom	σ	prom	σ	prom	σ
+FYG	0.14	0.06	0.10	0.06	0.12	0.09	0.13	0.07	0.16	0.09
+LVI	0.13	0.08	0.12	0.08	0.12	0.07	0.12	0.08	0.11	0.07
+RK	0.18	0.06	0.18	0.06	0.18	0.07	0.17	0.07	0.17	0.07
+RK-DE	0.12	0.09	0.12	0.08	0.11	0.09	0.11	0.10	0.11	0.09
+PST-RK	0.02	0.10	0.02	0.10	0.01	0.12	0.05	0.12	0.05	0.10

Abreviaciones:
prom = promedio
σ = desviación estándar

Tabla 10. Tablas de los valores obtenidos por IDDomainSpotter de las regiones predichas como ordenadas presentes en el N-terminal de Esg de *Dme* y sus ortólogos. En color verde se indica la composición de aminoácidos que se presenta con mayor abundancia utilizando un corte de 0.1.

Aminoácidos	Dme		Ami		Dre		Hsa		Hdu		Cin		Sko	
	S2 (116-148 aa)		S2 (20 -70 aa)		S2 (26-68 aa)		S2 (23-67 aa)		S2 (86-126 aa)		S2 (169-215 aa)		S2 (21-60 aa)	
	prom	σ	prom	σ	prom	σ	prom	σ	prom	σ	prom	σ	prom	σ
+FYG	0.18	0.08	0.12	0.06	0.06	0.05	0.14	0.04	0.09	0.06	0.23	0.11	0.11	0.09
+LVI	0.10	0.09	0.24	0.07	0.34	0.07	0.24	0.07	0.14	0.06	0.14	0.07	0.24	0.08
+RK	0.05	0.03	0.08	0.06	0.02	0.03	0.00	0.01	0.03	0.05	0.03	0.03	0.10	0.06
+RK-DE	0.02	0.04	0.02	0.06	-0.02	0.05	-0.06	0.05	0.00	0.03	0.02	0.03	0.00	0.09
+PST-RK	0.26	0.10	0.15	0.10	0.36	0.10	0.37	0.06	0.28	0.15	0.43	0.16	0.18	0.13

Aminoácidos	Oha		Xtr		Aqu		Aes		Spu			
	S2 (58-102 aa)		S2 (21-66 aa)		S2 (20-56 aa)		S2 (24-64 aa)		S2 (16-46 aa)		S4 (115-161 aa)	
	prom	σ	prom	σ	prom	σ	prom	σ	prom	σ	prom	σ
+FYG	0.11	0.06	0.15	0.05	0.26	0.09	0.12	0.06	0.16	0.11	0.14	0.07
+LVI	0.29	0.04	0.30	0.04	0.09	0.04	0.32	0.05	0.15	0.13	0.21	0.09
+RK	0.00	0.02	0.03	0.04	0.00	0.02	0.00	0.00	0.06	0.07	0.10	0.07
+RK-DE	-0.07	0.05	-0.05	0.05	-0.02	0.03	-0.07	0.05	-0.02	0.08	-0.01	0.06
+PST-RK	0.39	0.09	0.30	0.12	0.22	0.13	0.39	0.07	0.16	0.16	0.09	0.15

Aminoácidos	Fhe						Pvu			
	S2 (71-108 aa)		S4 (310-347 aa)		S6 (418-458 aa)		S2 (169-217 aa)		S4 (226-283 aa)	
	prom	σ	prom	σ	prom	σ	prom	σ	prom	σ
+FYG	0.19	0.05	0.16	0.07	0.15	0.05	0.21	0.11	0.14	0.06
+LVI	0.20	0.06	0.26	0.12	0.12	0.06	0.16	0.06	0.16	0.10
+RK	0.04	0.06	0.06	0.03	0.09	0.08	0.07	0.04	0.13	0.08
+RK-DE	-0.01	0.08	0.01	0.05	-0.03	0.16	0.03	0.05	0.01	0.07
+PST-RK	0.17	0.17	0.07	0.11	0.20	0.16	0.28	0.10	0.10	0.17

Abreviaciones:
prom = promedio
σ = desviación estándar

Para caracterizar la regiones propensas a ordenarse presentes en el N-terminal de Esg (Tabla 8) y conocer su composición y el tipo de IDR que representan, cada una de las secuencias de estas regiones se analizó con el servidor CIDER, cuyos gráficos se muestran en el Figura 12. De acuerdo a los resultados de los diagramas de estados, casi todas las regiones predichas como ordenadas presentes en el N-terminal de Esg de *Dme* y sus ortólogos caen en la región R1, la cual sugiere que pueden adoptar conformaciones compactas y esféricas de tipo glóbulo fundido, el cual corresponde a un estado de conformación compacta que puede contener elementos de estructura secundaria y es similar a un plegamiento terciario sin adoptar estructura tridimensional bien definida[5,16]. Sin embargo, la región denominada S3 de *Pvu*, está entre los límites de la región R1 y R2, lo cual indica que podría adoptar estructuras intermedias, es decir, estructuras compactas tipo glóbulo fundido y además otras extendidas.

Por otra parte, se ha reportado que las IDPs e IDRs muestran un perfil de composición caracterizado por una baja hidrofobicidad y una carga neta alta. Las gráficas de hidrofobicidad *vs*. carga neta mostraron que la mayoría de las regiones propensas a ordenarse presentes en el N-terminal de Esg de *Dme* (Figura 16) y sus ortólogos (Figura 13), presentan una hidrofobicidad y carga neta características de proteínas ordenadas. Sin embargo, las regiones S2 de *Hdu, Fhe* y *Pvu*, las regiones S4 de *Spu* y *Pvu* y la región S6 de Fhe presentan una hidrofobicidad y carga neta características de proteínas IDPs. Finalmente, las regiones S2 de *Sko* y S4 de *Fhe*, se encuentran entre el límite de composición de proteínas IDPs y ordenadas.

Figura 12. Diagramas de estados que muestran la clase conformacional que pueden adoptar las regiones predichas como ordenadas, presentes en el N-terminal de Esg de *Dme* y sus ortólogos, con base en su composición.

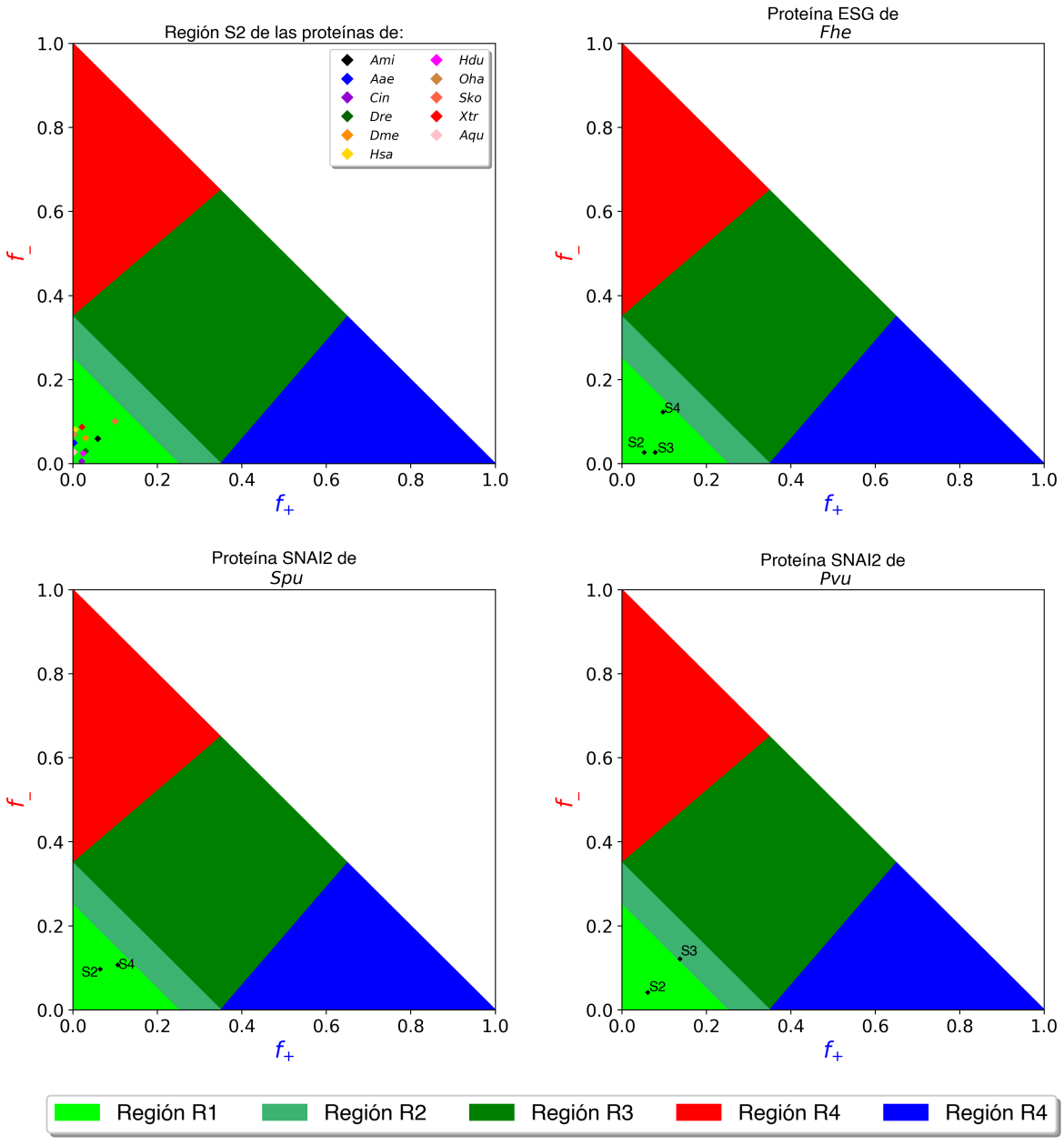
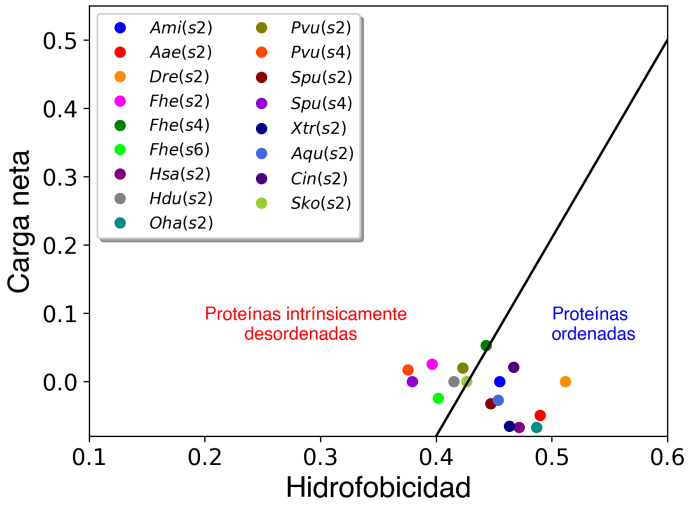


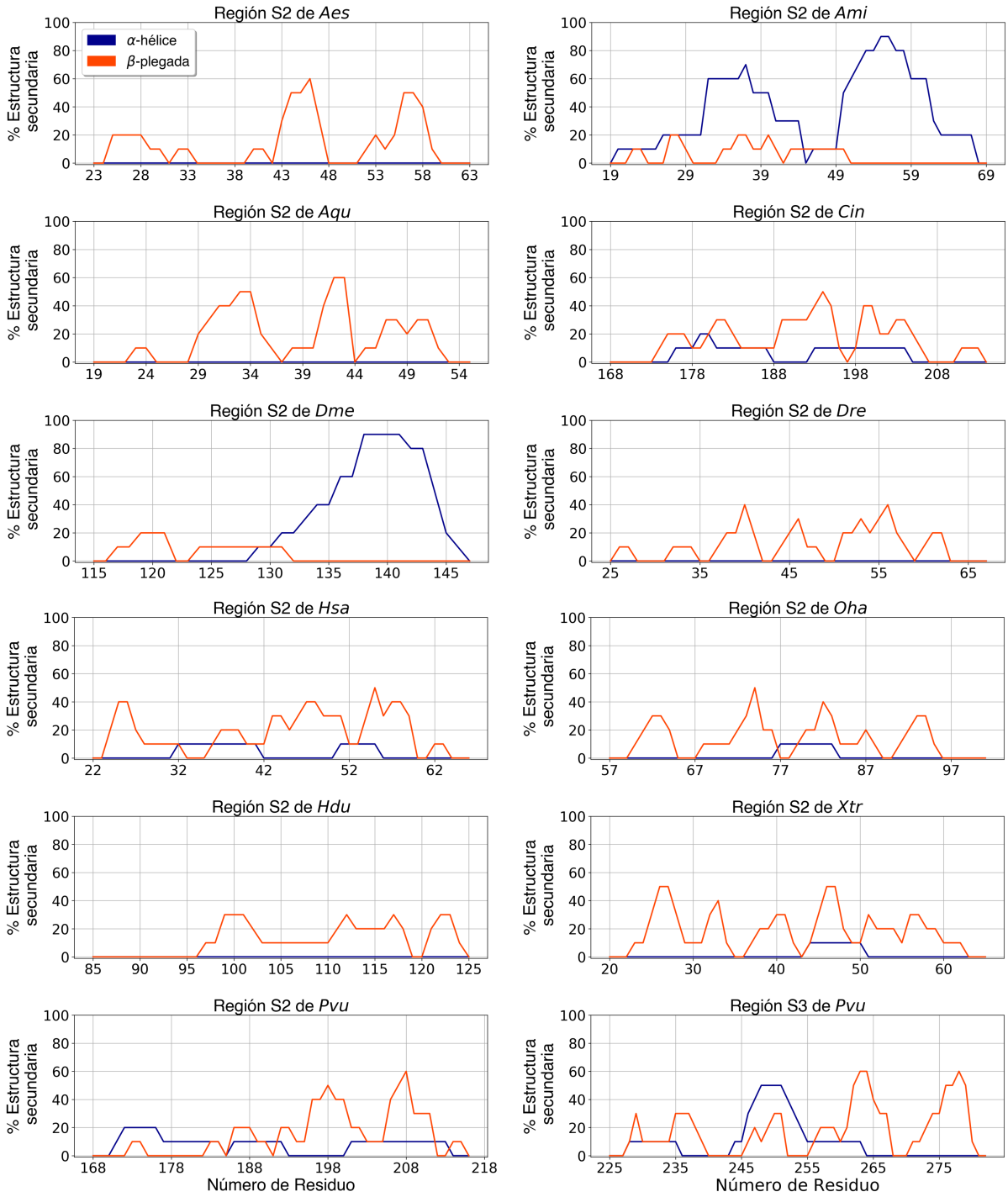
Figura 13. Gráfica de Hidrofobicidad y Carga neta de las regiones predichas como ordenadas, presentes en el N-terminal de Esg de *Dme* y sus ortólogos, con base en su composición. La recta negra separa las regiones de proteínas ordenadas y desordenadas.

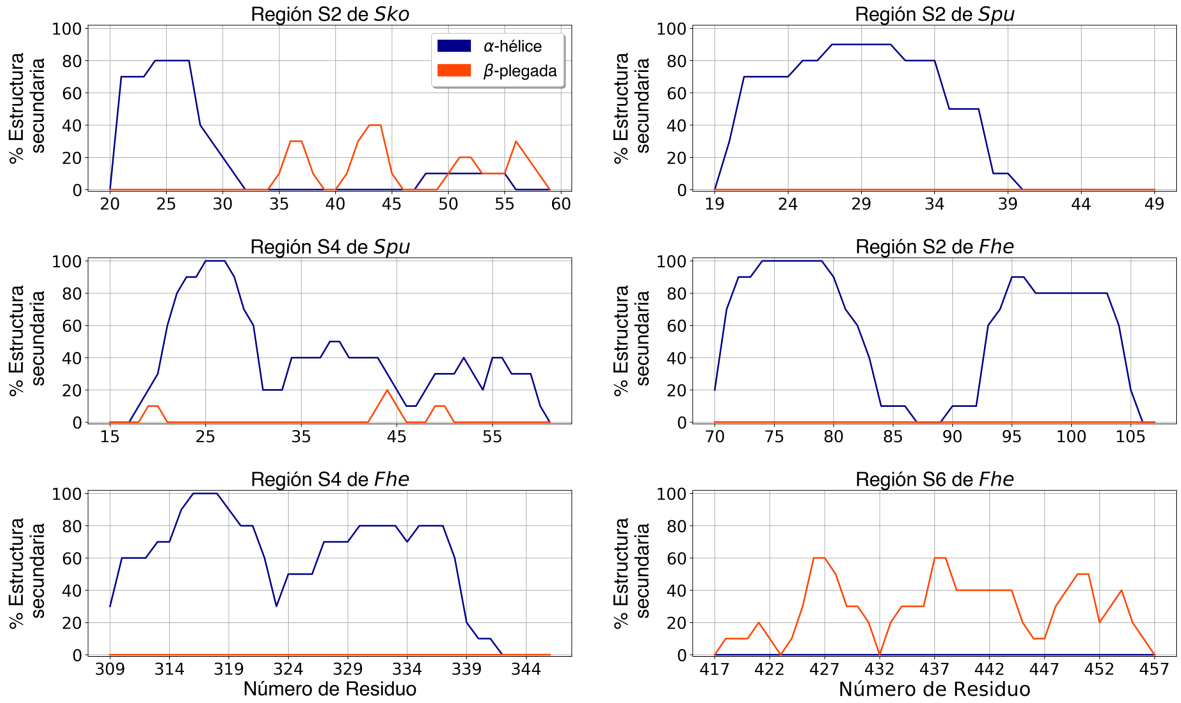


5.1.4 Análisis de predicción de estructura secundaria de regiones con tendencia a ordenarse

Para caracterizar el tipo de conformaciones que podrían adoptar las regiones predichas como regiones ordenadas, se utilizó el predictor de estructura SPARKS-X para generar modelos estructurales. Las gráficas de estructura secundaria se muestran en la Figura 14. Los resultados sugieren que la mayoría de las regiones predichas como ordenadas presentes en el N-terminal podrían adoptar conformaciones β-plegada. Por su parte, las regiones de *Ami, Dme, Sko*, y *Fhe* podrían adoptar estructura de α-hélice por arriba del 80%. El conjunto de resultados sugiere que las IDRs predichas como ordenadas presentes en el N-terminal de Esg de *Dme* y ortólogos pudieran ser IDRs de desorden flexible[2] ya que a nivel de secuencia no muestran un grado de conservación y la mayoría de los organismos estudiados muestra un perfil de desorden conservado. Consideramos de suma importancia y como perspectiva del análisis, caracterizar a nivel estructural las regiones predichas como ordenadas presentes en el N-terminal de Esg de *Dme* y sus ortólogos, las cuales pudieran ser probables β-MoRFs o α-MoRFs (Tabla 8) y a su vez, ser esenciales para ejercer sus funciones como factores de transcripción.

Figura 14. Predicción de estructura secundaria de las regiones predichas como ordenadas presentes en el N-terminal de Esg de *Dme* y sus ortólogos. En azul se representa el porcentaje de α-hélice, y en naranja el porcentaje de β-plegada.




Región S2 de *Sko*
α-hélice
β-plegada
% Estructura secundaria
Región S2 de *Spu*
Región S4 de *Spu*
Región S2 de *Fhe*
Región S4 de *Fhe*
Región S6 de *Fhe*
Número de Residuo

5.2 Perfil de desorden y análisis a nivel de secuencia de *Dme*-Esg

La importancia de detectar la presencia de IDRs y MoRFs en el N-terminal de *Dme*-Esg, radica en conocer y entender los mecanismos de reconocimiento molecular en los que participa este factor de transcripción. De acuerdo con el resultado del perfil de desorden del N-terminal de *Dme*-Esg (Figura 15), se muestra que el N-terminal es altamente desordenado a diferencia del C-terminal (residuos 310 al 470), el cual es estructurado y donde se encuentran los ZNFs. El N-terminal es una región desordenada y se dividió en tres principales regiones: S1 (residuos 1 al 110), S2 (residuos 111 al 155) y S3 (residuos 156 a 309), donde la región S2 tiene una mayor probabilidad a ordenarse a diferencia de la región S1 y S3, y la cual pudiera ser un MoRF.

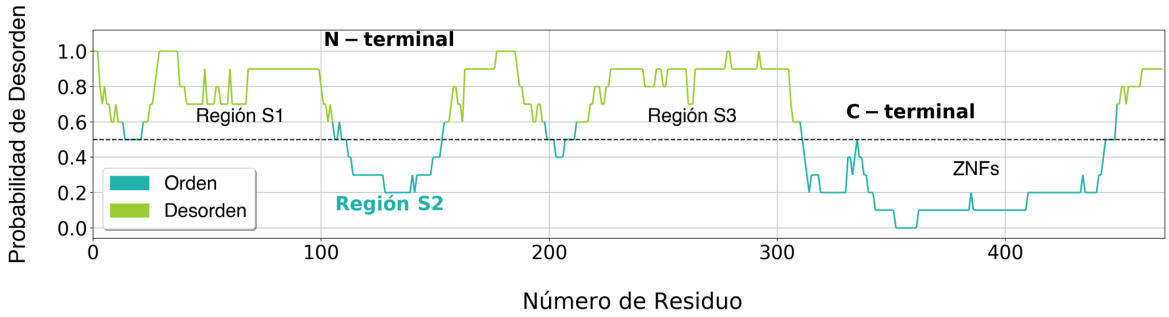


Figura 15. Perfil de desorden de la proteína *Dme*-Esg. La gráfica muestra la probabilidad de desorden por residuo. En color verde claro se muestran las regiones con probabilidad a mantener desorden, y en color verde fuerte se muestran las regiones con probabilidad a ordenarse.

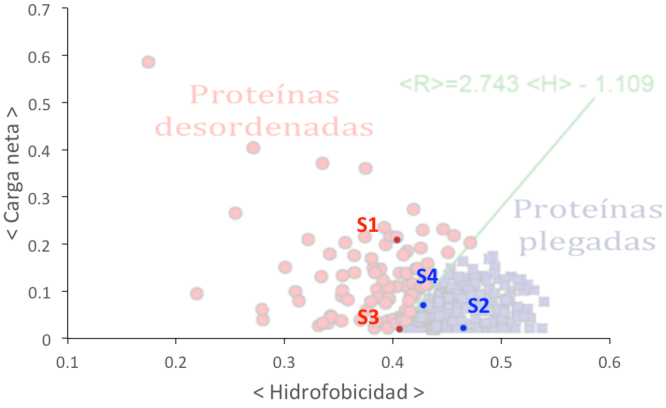


Figura 16. Gráfica de Hidrofobicidad vs Carga neta del promedio de cada una de las regiones de Esg. Las regiones desordenadas se muestran en rojo y las regiones ordenadas se muestran en azul.

Por su parte, considerando la gráfica de Uversky propuesta en el 2011 (Figura 3), se calculó el promedio de hidrofobicidad y carga neta en la escala[110] de Kyte y Doolittle

(1982) de cada una de las regiones de *Dme*-Esg (Figura 16). Los resultados sugieren que las regiones S1 y S3 presentan características de proteínas desordenadas, mientras que las regiones S2 y ZNFs presentan características de proteínas ordenadas.

5.3 Generación de conformaciones para la región S2 de *Dme-Esg* como estructuras iniciales para realizar dinámica molecular

La región S2 fue seleccionada para ser estudiada a través de simulaciones de MD. Un elemento indispensable en este proyecto son los modelos estructurales, los cuales son esenciales para iniciar las simulaciones de MD y poder estudiar su comportamiento estructural. Sin embargo, actualmente *Dme*-Esg no tiene homólogos estructurados y no hay estructuras de su N-terminal reportadas que hayan sido determinadas de manera experimental. Por lo anterior, predictores de estructura como HHpred, I-Tasser, Phyre2, QUARK y SPARKS-X fueron usados para predecir modelos estructurales de la región S2 (Figura 17). Las predicciones sugieren la presencia de un α-hélice en el C-terminal y algunas veces, la presencia de un α-hélice y β-hairpin en el N-terminal, excepto un modelo, el cual sugiere que la región S2 adopta estructura de β-plegada (Figura 17, SP6). Los resultados de estos predictores no fueron considerados como confiables tomando en cuenta su calificación en los predictores, debido en buena medida a la baja identidad de secuencia con dominios con estructura conocida (datos no mostrados), pero los modelos obtenidos fueron útiles como coordenadas iniciales para las simulaciones de MD.

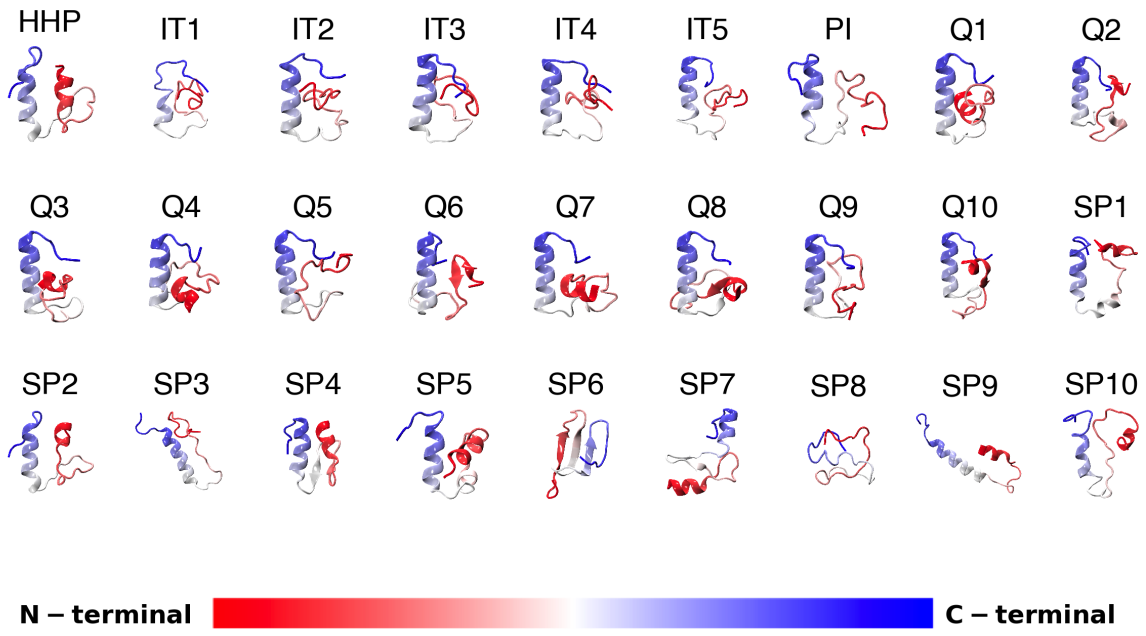


Figura 17. Ensamble inicial de la región S2 presente en el N-terminal de *Dme*-Esg obtenido por los predictores HHpred (HHP), I-Tasser (IT1 a IT5), Phyre2 (PI), QUARK (Q1 a Q10) y SPARKS-X (SP1 a SP10). Cada modelo muestra elementos de estructura secundaria y un código de color que va progresivamente de rojo a azul de N-terminal a C-terminal, respectivamente.

Las IDPs e IDRs son representadas como un ensamble dinámico, el cual es caracterizado por diferentes conformaciones. Las simulaciones de MD pueden ser usadas para generar muchas conformaciones de una IDP y caracterizar su ensamble conformacional a través de información dinámica y estructural[8,49]. De esta manera, no es un problema la calidad de las estructuras iniciales, porque a través de las simulaciones de MD las estructuras adoptan conformaciones que permiten capturar la dinámica conformacional de una proteína. Las transiciones conformacionales de una IDP o IDR pueden ocurrir en muchos ordenes de magnitud (desde ps a ms)[8]. En este trabajo, hemos considerado varias simulaciones cortas (2 μs) usando los 27 modelos generados en vez de una simulación larga con un solo modelo para explorar el ensamble conformacional de la región S2, debido a que esta estrategia permite un muestreo conformacional más eficiente e incrementa la probabilidad para converger con datos experimentales[111].

5.4 Validación de las simulaciones con CHARMM36 y solvente implícito

5.4.1 Simulaciones del péptido $(AAQAA)_3$

$(AAQAA)_3$ es un péptido helicoidal que ha sido bien estudiado y caracterizado experimentalmente para entender las transiciones hélice – coil[58,99,112]. Se ha reportado que $(AAQAA)_3$ tiene ~19% a 21% de contenido helicoidal[58,99] y ha sido usado como control para la optimización y parametrización de diferentes campos de fuerza. En este estudio, reportamos dos simulaciones de 16 μs de $(AAQAA)_3$, usando el mismo protocolo de simulación para la región S2 de Esg. Esta combinación en particular de CHARMM36 con el solvente GBSA no ha sido reportado como control para $(AAQAA)_3$, a pesar de haber sido utilizado para simular otras IDPs, como se muestra en la Tabla 1 y en el Tabla 11.

Tabla 11: Protocolos de simulación de MD de $(AAQAA)_3$ y su % de helicidad

Sistema	Solvente	Campo de fuerza	Tiempo total (acumulado)	% helicidad	Referencia
$(AAQAA)_3$	**Implícito (GBSW)**	**CHARMM22/CMAP**	**320 ns**	**~65%**	113
	Explícito (TIP3P)			**~90**	
	Explícito (TIP3P)	**ff03**	**960 ns para cada campo de fuerza**	**93.9%**	114
		ff99SB		**26.9%**	
		ff03*		**45.9%**	
		ff99SB*		**48.5%**	
	Explícito (TIP3P)	**CHARMM36**	**4.8 µs**	**~32%**	115
	Explícito (TIP3P)	**CHARMM36**	**4.8 µs**	**~44%**	32
	Implícito (EEF1-C19, EEF1-SB FACTS, SCPISM,)	**CHARMM36/EEF1-C9**	**100 ns para cada campo de fuerza**	**~12%**	
		CHARMM36/EEF1-SB		**~30%**	

		CHARMM36/FACTS		**~90%**	
		CHARMM36/SCPISM		**~85%**	
	Explícito (TIP3P)	**CHARMM36m (C36m)**	**16 μs para cada campo de fuerza**	**17%**	99
		CHARMM36 (C36)		**13%**	
	Implícito (GBMV2)	**CHARMM36**	**320 ns**	**42-47%**	112
	Explícito (TIP3P/TIP4P)	**C22*/TIP3P**	**20 μs para cada campo de fuerza**	**~30%**	116
		C36m/TIP3P a99SB-ILDN/TIP3P a03ws/TIP4P-D a99SB-ILDN/TIP4P-D a99SB/TIP4P-Ew		**5-12%**	
	Explícito (TIP3P modificada)	**C36IDPSFF**	**5 μs**	**~10%**	117

Nota: ff99SB* y ff03* son campos de fuerza con correcciones en φ

Una de las simulaciones comenzó de una conformación extendida (0% helicoidal), y la otra comenzó de una conformación 100% helicoidal, lo que explica el 50% de helicidad por residuo para las estructuras iniciales (Figura 18B). La Figura 18 muestra que $(AAQAA)_3$ intercambia frecuentemente conformaciones coil y helicoidales (Figuras 18A y 18C), y su contenido helicoidal promedio durante los 32 μs fue de ~24%, lo que indica que CHARMM36 con GBSA proporciona un equilibrio razonable entre las conformaciones de hélice y coil para $(AAQAA)_3$. Así mismo, se observa en la Figura 18A que las dos simulaciones de 16 μs presentan un comportamiento diferente: una muestra hélice con periodos largos de ausencia de hélices (simulación 1) y la otra, un exceso de hélice (simulación 2). Esto indica que incluso para sistemas simples como este pequeño péptido, haber visto varias veces la transición entre hélice - coil no garantiza un muestreo suficiente.

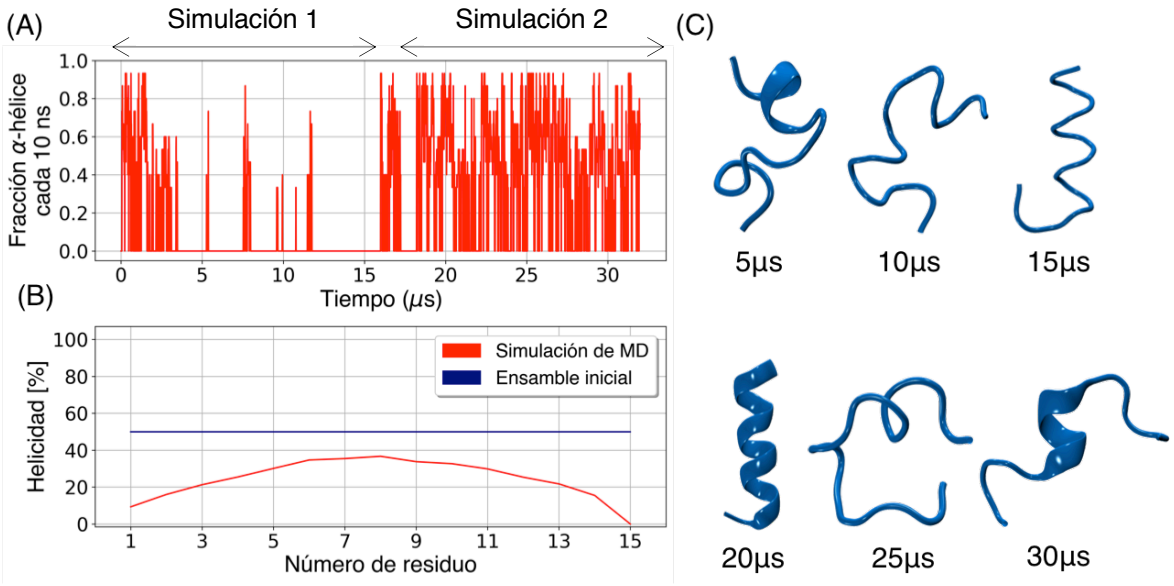


Figura 18. Ensamble de $(AAQAA)_3$ durante 32 μs of MD simulación. (A) Fracción de α-hélice de $(AAQAA)_3$ calculado en bloques de 10ns; los primeros 16 μs corresponden a la simulación que comenzó de la conformación extendida (simulación 1) y los últimos 16 μs corresponden a la simulación que comenzó con la conformacional helicoidal (simulación 2). (B) Porcentaje de helicidad por residuo, promedio sobre los 32 μs (línea roja) comparada con el ensamble inicial (línea azul). (C) Conformaciones de $(AAQAA)_3$ que representan las transiciones de hélice - coil en diferentes tiempos durante las simulaciones.

Para explorar la posibilidad de la formación de un núcleo hidrofóbico, calculamos el mapa de contactos para el ensamble de 32 μs de simulación de $(AAQAA)_3$, el cual se muestra en el Figura 19. Las estructuras que representan las interacciones más comunes se incluyen en el Figura 20 y, como se esperaba debido a su mayor número de átomos de carbono, involucran los residuos de Glutamina. Es claro, a partir de este mapa de contactos, que no se promueven contactos de largo alcance en este péptido, que es 80% hidrófobo.

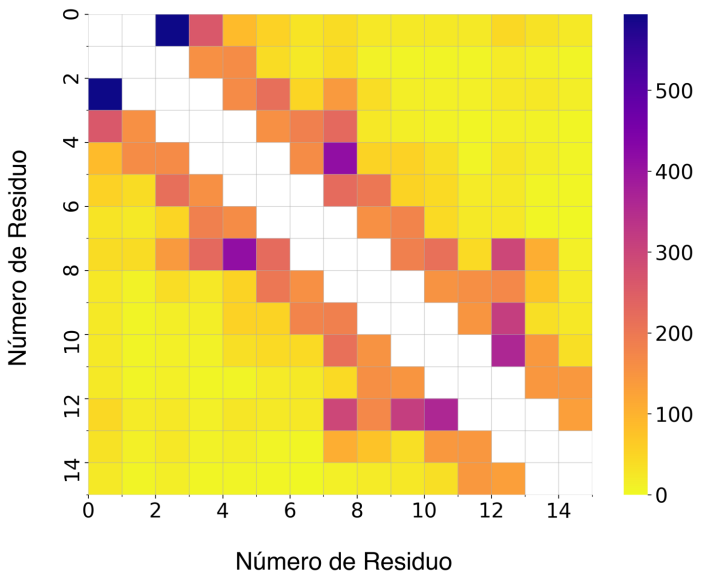


Figura 19: Mapa de calor (interacciones entre átomos carbono – carbono) dentro de una distancia de 6Å durante 32 μs de simulación de $(AAQAA)_3$. Los contactos entre residuos que son vecinos inmediatos no son considerados.

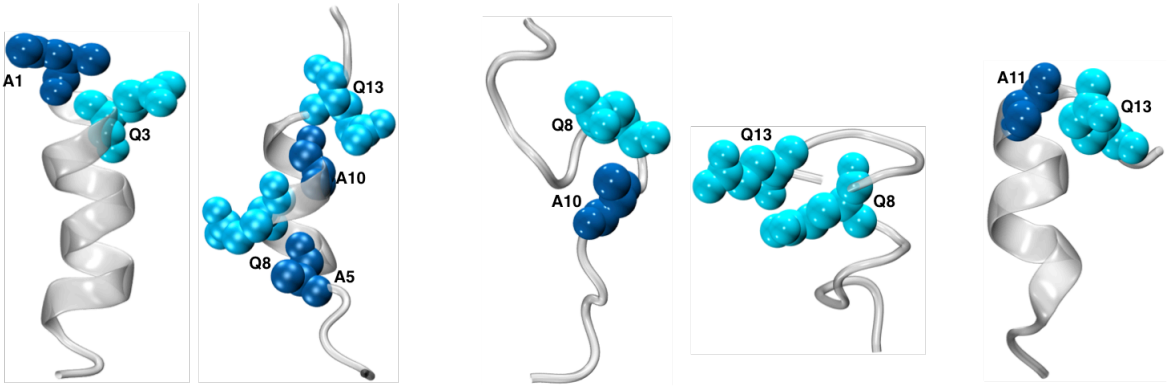


Figura 20: Las interacciones entre residuos más frecuentes en el mapa de contactos (Figura 19) de $(AAQAA)_3$. El trazo de cadena principal de la proteína se muestra en gris, y los residuos de Alanina y Glutamina como esferas de van der Waals (azul y cyan, respectivamente).

5.4.2 Análisis de la población α_L

Es bien conocido que en las simulaciones de IDPs se presenta un incremento artificial de estructura secundaria y terciaria. Una versión modificada de CHARMM36 (CHARMM36m) corrige el exceso de población de α-hélice levógira (α_L) generada en simulaciones con CHARMM36[99]. Es importante mencionar que esta mejora no estaba disponible cuando comenzamos este proyecto, por lo que todas las simulaciones se llevaron a cabo con el mismo campo de fuerza para mantener la consistencia. Para determinar el nivel de inexactitud de nuestra descripción de S2, calculamos la población de α_L durante los 54 μs de simulación del ensamble de la región S2, la cual resultó ser de 1.36%, y se encontró que es significativamente baja en comparación con otras simulaciones de IDPs generadas con CHARMM36 (entre 5.7% a 20%) y considerada dentro del margen de error de los datos experimentales[58,99].

5.4.3 Monitores de convergencia estructural

Para caracterizar estructuralmente a la región S2 presente en el N-terminal de *Dme*-Esg y conocer si el tiempo de simulación ha sido suficiente para comenzar a observar convergencia estructural, calculamos la raíz de la desviación cuadrática media (RMSD) y radio de giro (Rg) para cada estructura en el ensamble simulado. RMSD es una medida de similitud estructural y flexibilidad[37] y fue calculada considerando la desviación de los átomos de C_α de las estructuras iniciales con respecto a cada una de las estructuras obtenidas en las trayectorias de la simulación de MD, con el objetivo de determinar cuánto habían cambiado estructuralmente en 2 μs; por otro lado, Rg es útil para describir la compactación estructural[35]. Valores bajos de RMSD indican que hay una alta similitud entre conformaciones, y valores bajos de Rg indican que hay una mayor compactación. El paisaje energético construido con las variables RMSD y Rg (Figura 21A) durante los 54 μs de simulación, muestra que la distribución conformacional está localizada en dos principales poblaciones, una población pequeña con valores bajos tanto de RMSD (entre 0 Å a 2 Å) y Rg (entre 10 Å a 11 Å), así como una población más grande, con variación en los valores de RMSD (entre 5 Å a 11 Å) y Rg (entre 9.5 Å a 12 Å). Tener estructuras que presentan valores bajos de RMSD y a su vez valores bajos de Rg, podría ser un indicador de que los contactos intramoleculares estén dificultando la exploración conformacional.

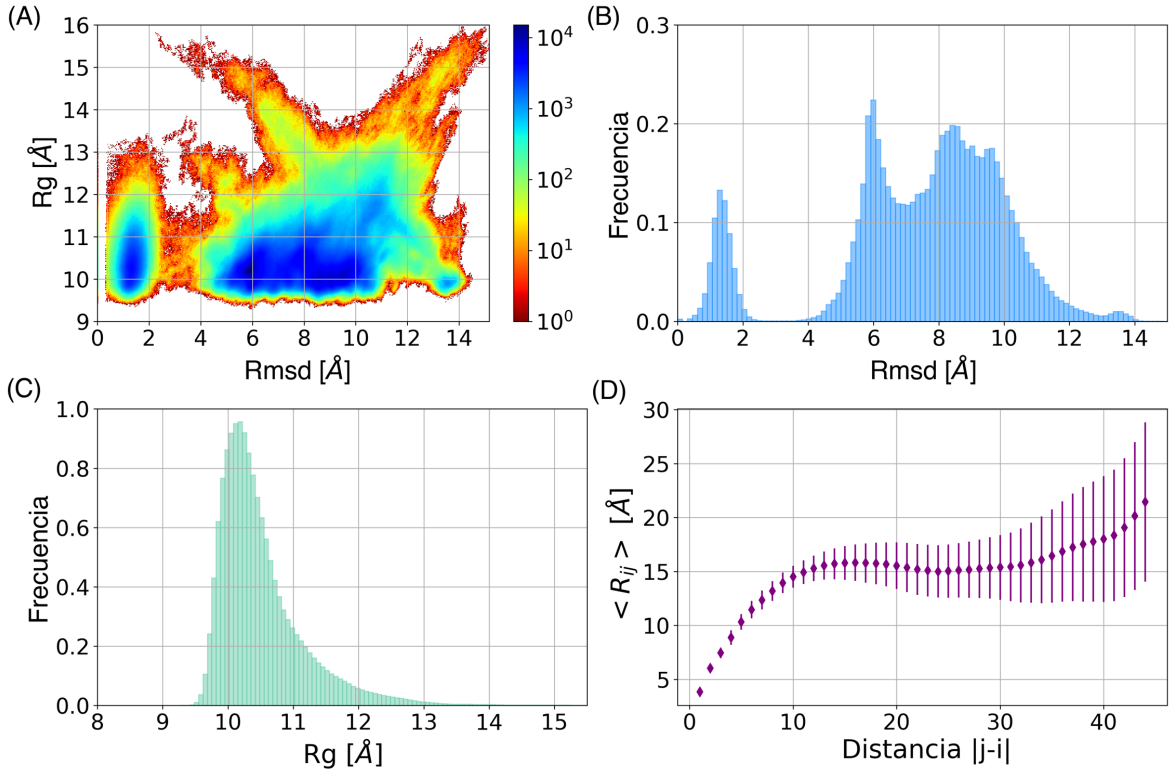


Figura 21. Diversidad estructural y grado de compactación de la región S2 de *Dme*-Esg. (A) Paisaje energético construido con las variables RMSD [Å] y Rg [Å]. (B) Histograma de RMSD [Å] calculado a partir de la desviación de los átomos de C_α en el ensamble respecto a las estructuras iniciales durante 54 μs de simulación. (C) Histograma de Rg [Å] calculado para el ensamble de 54 μs de simulación. (D) Promedio de las distancias entre residuos [Å] en función de la distancia en secuencia durante 54 μs de simulación.

Los histogramas de RMSD y Rg se muestran en las Figuras 21B y 21C, respectivamente. La distribución de RMSD mostró cinco picos localizados cerca de 1.5 Å, 6 Å, 8.5 Å, 9.5 Å y 13.5 Å, los cuales indican que hay conformaciones que se mantuvieron muy cerca de su punto inicial, mientras que otras son estructuralmente diferentes (Figura 21B). Considerando las propiedades de secuencia de IDPs, es posible estimar el radio hidrodinámico (Rh)[118]. Para la región S2 con 45 residuos, el Rh estimado de manera teórica presenta un valor alrededor de 15 Å, y considerando que el diámetro de una molécula de agua es ~3 Å, el Rg debería ser ~12 Å. La distribución de Rg mostró un solo pico localizado entre 10 Å y 11 Å, el cual corresponde a conformaciones semicompactas (Figura 21C). Los resultados mostraron que el ensamble de la región S2 podría adoptar

conformaciones heterogéneas y flexibles, características de una IDP[49]. Algunos campos de fuerza y modelos de solvente implícito generan una sobreestimación de compactación para los ensambles de IDPs[39,119], lo cual se pudo observar en nuestras simulaciones al presentar un pequeño aumento en la compactación. Sin embargo, la compactación de una IDP o IDR depende de su secuencia, por ejemplo, la fracción de residuos cargados y el contenido de prolina[43,120]. El Rg y Rh se han relacionado con la carga neta por residuo (NCPR)[43], donde las IDPs con NCPR>0.25 adoptan conformaciones extendidas tipo coil, mientras que una NCPR<0.25 indica conformaciones globulares y compactas. La región S2 de *Dme*-Esg tiene tres residuos cargados, una carga neta de +1, 21 residuos hidrofóbicos, 14 residuos polares, 7 residuos aromáticos y 6 Prolinas; además 22 de sus 45 aminoácidos son promotores de desorden. La región S2 de Esg tiene un NCPR=0.022 obtenido con el servidor CIDER[84], por lo que se espera que adopte conformaciones compactas.

La Figura 21D muestra que el promedio de las distancias entre residuos son mayores que las esperadas para una estructura colapsada de Lennard-Jones[43,119], pero mucho más pequeñas que los de una cadena Flory de la misma longitud. Es de suma importancia considerar que no hay datos experimentales para S2 que podamos usar para guiar nuestras simulaciones o cómo prueba de la calidad del conjunto de datos.

Uno de los principales objetivos en las simulaciones de MD de IDPs es mostrar convergencia entre conformaciones. Una de las ventajas de generar múltiples simulaciones de manera independiente y a partir de diferentes estructuras, es acelerar la convergencia estructural. Para determinar si 54 μs de simulación es tiempo suficiente para observar convergencia estructural, se buscaron estructuras que fueran estructuralmente similares en distintas simulaciones. Para ello se construyó una matriz de RMSD 2D comparando dos diferentes trayectorias de MD; de esta manera, cada estructura generada se compara con el resto de las otras.

La Figura 22 muestra los mapas de calor con la comparación de la RMSD calculada entre átomos de C_α de las estructuras generadas de las trayectorias de los modelos SP8 y SP9, y las trayectorias de los modelos IT4 y SP7. Los pares de estructuras con la RMSD más pequeña se encuentran en azul, mientras que la RMSD más grande se encuentra en color rojo. La línea diagonal azul representa la comparación de la RMSD de una estructura con

ella misma y los dos cuadros diagonales (números de instantánea 1 - 2000 y 2001 - 4000) corresponden a la RMSD entre estructuras de la misma trayectoria. La parte interesante de estos gráficos corresponde a los cuadros fuera de la diagonal, donde se comparan las estructuras entre trayectorias. La RMSD por modelo muestra la transición entre los estados conformacionales de cada trayectoria.

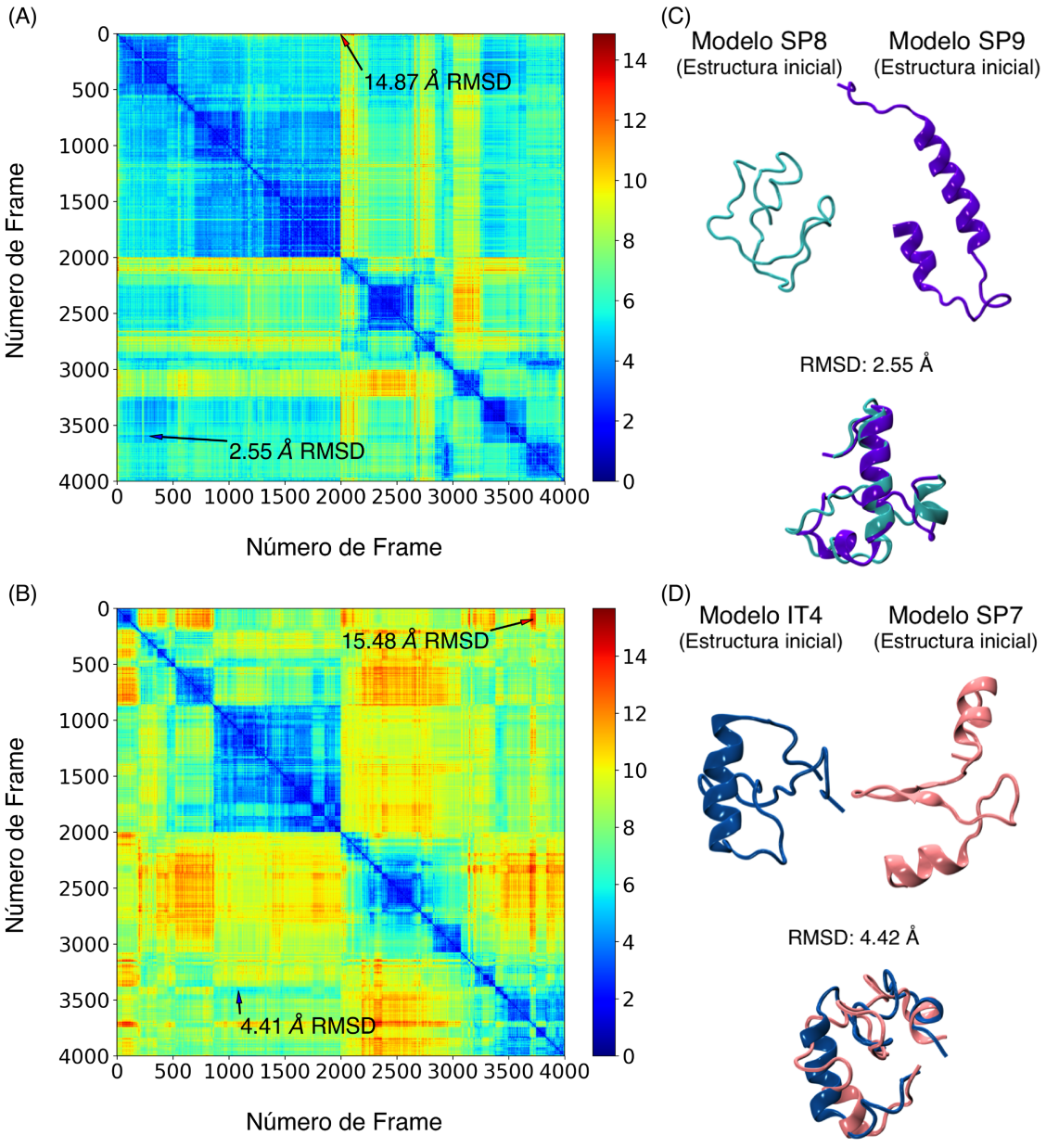


Figura 22. Mapas de calor que representan la RMSD [Å] entre pares de estructuras calculada sobre átomos de C_α de las trayectorias de los modelos (A) SP8 (número de instantáneas 1 – 2000) y SP9 (número de instantáneas 2001 – 4000) e (B) IT4 (número de instantáneas 1 –

2000) y SP7 (número de instantáneas 2001 - 4000). Cada gráfica muestra la RMSD mínima y máxima por pares de estructuras. (C) Estructuras iniciales de los modelos SP8 y SP9, y alineamiento estructural de los modelos que mostraron la mínima RMSD entre pares de estructuras. (D) Estructuras iniciales de los modelos IT4 y SP7, y alineamiento estructural de la RMSD mínima entre pares de estructuras.

En la Figura 22A, se muestra la matriz RMSD 2D de la comparación de las trayectorias de los modelos SP8 y SP9. La RMSD más pequeña fue 2.55 Å, y la máxima RMSD fue de 14.87 Å; así mismo se observan valores bajos de RMSD que ocurren varias veces durante estas dos trayectorias. La comparación de las estructuras iniciales de los modelos SP8 y SP9 mostró una RMSD de 14.1 Å (Tabla 12), y durante 2 μs de simulación de cada una, las trayectorias mostraron la distancia más pequeña y más grande del resto de los modelos (Tabla 13). Resultados similares fueron obtenidos con las trayectorias de los modelos IT4 y SP7, donde la RMSD menor fue de 4.42 Å, y la máxima RMSD fue de 15.48 Å, lo cual se puede observar en la Figura 22C. La comparación de las estructuras iniciales de los modelos IT4 y SP7 mostraron una RMSD de 10.9 Å (Tabla 12), y durante 2 μs de simulación de cada uno, las trayectorias mostraron la mayor RMSD de los valores mínimos RMSD del resto de los modelos (Tabla 13). El alineamiento estructural de las dos estructuras con la mínima y máxima RMSD es mostrado en las Figuras 22C y 22D, ilustrando el grado de diversidad conformacional de cada modelo. Las diferencias conformacionales pueden ser atribuidas a los movimientos de la cadena principal para adoptar estructura secundaria o no. En proteínas plegadas, un corte de ~ 2 Å de RMSD es suficiente para considerar similitud estructural; en IDPs un corte de ~ 2.55 Å a 4.42 Å de RMSD podría ser suficiente para considerar similitud estructural. Estos resultados sugieren que simular 2 μs cada modelo, es tiempo razonable y suficiente para ver convergencia en el muestreo conformacional de la región S2 y para comenzar a describir los estados conformacionales accesibles a ella. En este estudio, se asume que las estructuras generadas por los 54 μs de simulación de MD pueden representar la diversidad del espacio conformacional de la región S2 de *Dme*-Esg.

Tabla 12. Comparación de la distancia de RMSD por pares de las estructuras iniciales de la región S2 de *Dme*-Esg. Los cuadros de color rojo y azul indican la distancia menor y mayor de RMSD por pares, respectivamente.

Matriz de la distancia RMSD por pares (en Å) entre Cα de las estructuras iniciales de la región S2 de Esg.																										
	HHP	IT1	IT2	IT3	IT4	IT5	PI	Q1	Q10	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	SP1	SP10	SP2	SP3	SP4	SP5	SP6	SP7	SP8	SP9
HHP		9.5	8.48	9.6	8.6	8.5	9.6	9.1	7.5	5.8	9.3	7.5	7.76	5.9	7.0	7.8	8.7	8.4	9.8	3.0	10.4	3.8	7.8	9.8	10.9	8.2	11.8
IT1	9.5		3.70	3.0	4.6	5.5	9.0	9.2	6.3	8.3	8.6	6.8	8.11	8.3	6.3	5.2	5.4	7.4	9.1	9.9	10.1	9.7	7.1	8.7	10.9	7.2	12.3
IT2	8.5	3.7		3.7	3.9	6.1	8.8	9.1	5.7	7.0	8.6	6.4	7.6	7.2	5.4	4.8	6.0	6.8	8.7	8.6	10.3	8.1	7.4	8.8	11.0	6.8	12.2
IT3	9.6	3.0	3.7		4.6	5.8	8.6	9.1	6.2	8.3	8.2	7.4	7.8	8.7	6.8	4.9	6.3	7.1	8.9	9.9	10.6	9.6	7.3	8.5	10.9	7.8	12.1
IT4	8.6	4.6	3.9	4.6		5.5	8.4	9.6	6.5	7.0	8.5	85.8	8.0	6.9	5.4	4.7	5.3	7.4	9.2	8.6	11.1	8.4	7.7	8.2	10.9	6.7	12.6
IT5	8.5	5.5	6.1	5.8	5.5		8.1	9.3	7.8	8.6	8.3	6.5	8.7	7.8	5.6	6.7	4.8	7.7	8.4	8.9	9.8	8.8	6.4	9.1	10.4	8.2	11.6
PI	9.6	9.0	8.8	8.6	8.4	8.1		9.4	9.4	10.2	10.6	9.1	10.4	9.6	8.7	7.5	8.1	8.5	8.1	10.3	11.0	9.7	8.0	8.2	7.8	10.2	11.9
Q1	9.1	9.2	9.1	9.1	9.6	9.3	9.4		8.2	8.0	4.1	9.0	6.1	9.1	8.7	8.6	8.7	8.5	9.1	9.6	8.6	8.9	8.4	7.8	7.2	8.8	13.0
Q10	7.5	6.3	5.7	6.2	6.5	7.8	9.4	8.2		5.1	7.4	7.5	6.0	6.9	6.8	6.3	7.5	7.4	9.6	7.2	10.5	6.8	8.6	9.8	9.8	8.3	10.4
Q2	5.8	8.3	7.0	8.3	7.0	8.6	10.2	8.0	5.1		7.2	7.0	4.6	4.1	6.5	6.2	8.5	7.8	9.6	5.4	9.2	4.9	8.7	10.2	9.6	7.8	11.3
Q3	9.3	8.6	8.6	8.2	8.5	8.3	10.6	4.1	7.4	7.2		8.6	5.5	8.4	8.5	7.9	8.6	9.5	10.0	9.4	8.9	8.4	8.4	8.2	8.0	9.1	12.5
Q4	7.5	6.8	6.4	7.4	5.8	6.5	9.1	9.0	7.5	7.0	8.6		8.8	4.8	4.6	6.5	4.1	7.9	8.7	8.4	10.7	8.1	6.0	8.1	10.9	6.3	12.5
Q5	7.8	8.1	7.6	7.8	8.0	8.7	10.4	6.1	6.0	4.6	5.5	8.8		6.2	8.1	6.8	8.9	7.7	8.7	7.3	7.8	6.6	9.5	9.8	8.1	8.8	12.2
Q6	5.9	8.3	7.2	8.7	6.9	7.8	9.6	9.1	6.9	4.1	8.4	4.8	6.2		5.3	6.3	6.9	7.5	9.3	6.5	9.4	6.0	7.3	9.5	10.7	7.0	12.1
Q7	7.0	6.3	5.4	6.8	5.4	5.6	8.7	8.7	6.8	6.5	8.5	4.6	8.1	5.3		6.3	5.3	7.0	8.4	7.1	10.4	7.1	6.3	8.4	9.9	7.9	11.9
Q8	7.8	5.2	4.8	4.9	4.7	6.7	7.5	8.6	6.3	6.2	7.9	6.5	6.8	6.3	6.3		5.8	8.6	10.6	8.2	11.2	7.6	7.9	7.3	9.1	6.1	11.9
Q9	8.7	5.4	6.0	6.3	5.3	4.8	8.1	8.7	7.5	8.5	8.6	4.1	8.9	6.9	5.3	5.8		8.3	9.1	9.6	10.6	9.4	5.9	7.9	9.6	6.7	12.7
SP1	8.4	7.4	6.8	7.1	7.4	7.7	8.5	8.5	7.4	7.8	9.5	7.9	7.7	7.5	7.0	8.6	8.3		6.1	8.6	8.0	8.5	6.5	11.1	9.9	9.3	11.8
SP10	9.8	9.1	8.7	8.9	9.2	8.4	8.1	9.1	9.6	9.6	10.0	8.7	8.7	9.3	8.4	10.6	9.1	6.1		9.8	6.7	9.8	7.8	10.4	8.5	10.5	13.0
SP2	3.0	9.9	8.6	9.9	8.6	8.9	10.3	9.6	7.2	5.4	9.4	8.4	7.3	6.5	7.1	8.2	9.6	8.6	9.8		10.1	3.3	8.6	10.6	11.0	8.8	12.0
SP3	10.4	10.1	10.3	10.6	11.1	9.8	11.0	8.6	10.5	9.2	8.9	10.7	7.8	9.4	10.4	11.2	10.6	8.0	6.7	10.1		10.1	9.5	12.9	10.4	11.0	12.4
SP4	3.8	9.7	8.1	9.6	8.4	8.8	9.7	8.9	6.8	4.9	8.4	8.1	6.6	6.0	7.1	7.6	9.4	8.5	9.8	3.3	10.1		8.4	10.2	10.2	8.0	12.3
SP5	7.8	7.1	7.4	7.3	7.7	6.4	8.0	8.4	8.6	8.7	8.4	6.0	9.5	7.3	6.3	7.9	5.9	6.5	7.8	8.6	9.5	8.4		8.4	10.3	8.2	11.3

SP6	9.8	8.7	8.8	8.5	8.2	9.1	8.2	7.8	9.8	10.2	8.2	8.1	9.8	9.5	8.4	7.3	7.9	11.1	10.4	10.6	12.9	10.2	8.4		7.4	8.8	13.9
SP7	10.9	10.9	11.0	10.9	10.9	10.4	7.8	7.2	9.8	9.6	8.0	10.9	8.1	10.7	9.9	9.1	9.6	9.9	8.5	11.0	10.4	10.2	10.3	7.4		10.9	12.6
SP8	8.2	7.2	6.8	7.8	6.7	8.2	10.2	8.8	8.3	7.8	9.1	6.1	8.8	7.0	7.9	6.1	6.7	9.3	10.5	8.8	11.0	8.0	8.2	8.8	10.9		14.1
SP9	11.8	12.3	12.2	12.1	12.6	11.6	11.9	13.0	10.4	11.3	12.5	12.5	12.2	12.1	11.9	12.0	12.7	11.8	13.0	12.0	12.4	12.3	11.3	13.9	12.6	14.1	

Tabla 13. La distancia mínima del RMSD por pares durante 2 μs de simulación de cada modelo de la región S2 de *Dme*-Esg, comparada con las estructuras generadas por las simulaciones de los otros modelos. El cuadro naranja indica la distancia más baja de RMSD por pares entre cada par de simulaciones. El cuadro verde y amarillo indican la distancia más pequeña y la distancia más grande de las distancias mínimas de RMSD por pares, respectivamente.

Matriz de la distancia RMSD por pares (en Å) entre Cα durante 2 μs de simulación de cada estructura de la región S2 e Esg.																									
	HHP	IT1	IT2	IT3	IT4	IT5	PI	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	SP1	SP2	SP3	SP4	SP5	SP6	SP7	SP8	SP9
HHP		5.19	5.96	7.23	6.24	6.78	6.36	7.85	4.36	5.46	4.94	5.36	5.18	4.39	5.94	4.35	5.87	6.22	3.77	5.33	4.20	5.18	6.34	6.20	5.32	5.60
IT1	5.19		3.57	3.69	4.10	5.56	6.01	6.20	5.99	6.39	4.41	5.61	4.47	3.76	3.84	3.68	6.56	5.78	6.23	5.58	5.35	3.63	5.69	6.92	5.01	5.12
IT2	5.96	3.57		3.36	3.68	4.18	5.47	6.69	5.64	6.38	3.89	5.46	3.80	3.74	2.98	4.61	5.09	5.58	5.69	5.58	5.40	3.59	4.65	6.10	3.21	3.00
IT3	7.23	3.69	3.36		4.38	5.09	5.80	7.04	5.36	6.01	4.50	5.54	3.67	3.86	3.11	4.94	5.97	5.19	6.44	5.23	5.71	4.19	5.11	6.39	4.35	4.03
IT4	6.24	4.10	3.68	4.38		2.65	4.66	5.96	5.79	6.64	5.12	5.88	4.76	4.66	3.31	5.71	4.27	5.85	6.52	6.09	5.43	4.14	4.95	4.42	4.89	4.75
IT5	6.78	5.56	4.18	5.09	2.65		3.67	4.82	5.76	5.44	4.44	5.21	5.05	4.30	4.72	5.49	4.74	5.46	5.40	4.90	5.73	4.99	4.86	4.98	4.95	4.43
PI	6.36	6.01	5.47	5.80	4.66	3.67		4.23	4.86	3.68	5.61	3.83	5.44	5.53	5.66	7.07	3.62	3.87	4.06	3.90	5.20	6.64	5.80	3.75	6.30	6.08
Q1	7.85	6.20	6.69	7.04	5.96	4.82	4.23		5.13	3.65	6.87	6.03	6.54	6.10	6.18	6.92	3.81	5.03	4.80	3.43	6.04	5.15	6.46	5.60	6.58	6.96
Q2	4.36	5.99	5.64	5.36	5.79	5.76	4.86	5.13		3.30	5.03	4.22	5.07	4.41	5.97	6.11	5.45	5.10	3.59	4.11	3.52	6.29	6.15	5.35	6.10	5.35
Q3	5.46	6.39	6.38	6.01	6.64	5.44	3.68	3.65	3.30		5.75	3.21	5.94	5.22	5.77	7.00	3.82	5.15	3.88	3.24	4.25	6.78	6.64	5.66	6.28	5.83
Q4	4.94	4.41	3.89	4.50	5.12	4.44	5.61	6.87	5.03	5.75		4.03	2.70	2.86	4.36	3.70	6.79	5.36	6.35	5.86	3.52	5.00	5.77	6.22	3.60	3.59
Q5	5.36	5.61	5.46	5.54	5.88	5.21	3.83	6.03	4.22	3.21	4.03		3.57	3.48	5.46	5.51	6.53	3.89	4.68	4.06	3.98	6.33	6.05	5.50	5.48	4.58
Q6	5.18	4.47	3.80	3.67	4.76	5.05	5.44	6.54	5.07	5.94	2.70	3.57		3.01	3.83	4.00	5.14	4.86	5.37	5.14	4.07	3.75	5.64	5.20	3.88	3.80
Q7	4.39	3.76	3.74	3.86	4.66	4.30	5.53	6.10	4.41	5.22	2.86	3.48	3.01		4.16	4.50	5.69	5.65	6.00	4.85	3.82	4.46	5.24	5.70	4.18	3.97
Q8	5.94	3.84	2.98	3.11	3.31	4.72	5.66	6.18	5.97	5.77	4.36	5.46	3.83	4.16		4.99	5.37	5.68	6.02	6.05	5.23	3.67	5.17	5.72	3.30	3.56
Q9	4.35	3.68	4.61	4.94	5.71	5.49	7.07	6.92	6.11	7.00	3.70	5.51	4.00	4.50	4.99		6.77	7.43	7.39	6.88	4.49	3.93	5.98	6.78	3.80	4.26
Q10	5.87	6.56	5.09	5.97	4.27	4.74	3.62	3.81	5.45	3.82	6.79	6.53	5.14	5.69	5.37	6.77		4.71	4.82	4.60	6.28	4.58	6.50	4.84	6.06	5.66
SP1	6.22	5.78	5.58	5.19	5.85	5.46	3.87	5.03	5.10	5.15	5.36	3.89	4.86	5.65	5.68	7.43	4.71		4.77	3.82	5.60	6.55	6.28	6.02	6.42	6.00
SP2	3.77	6.23	5.69	6.44	6.52	5.40	4.06	4.80	3.59	3.88	6.35	4.68	5.37	6.00	6.02	7.39	4.82	4.77		2.87	3.43	6.14	6.64	5.68	6.02	4.86
SP3	5.33	5.58	5.58	5.23	6.09	4.90	3.90	3.43	4.11	3.24	5.86	4.06	5.14	4.85	6.05	6.88	4.60	3.82	2.87		4.35	5.74	5.78	5.35	5.44	5.03
SP4	4.20	5.35	5.40	5.71	5.43	5.73	5.20	6.04	3.52	4.25	3.52	3.98	4.07	3.82	5.23	4.49	6.28	5.60	3.43	4.35		5.73	6.29	5.98	3.81	4.47

SP5	5.18	3.63	3.59	4.19	4.14	4.99	6.64	5.15	6.29	6.78	5.00	6.33	3.75	4.46	3.67	3.93	4.58	6.55	6.14	5.74	5.73		4.41	6.06	4.01	3.89
SP6	6.34	5.69	4.65	5.11	4.95	4.86	5.80	6.46	6.15	6.64	5.77	6.05	5.64	5.24	5.17	5.98	6.50	6.28	6.64	5.78	6.29	4.41		5.65	5.86	5.47
SP7	6.20	6.92	6.10	6.39	4.42	4.98	3.75	5.60	5.35	5.66	6.22	5.50	5.20	5.70	5.72	6.78	4.84	6.02	5.68	5.35	5.98	6.06	5.65		5.39	6.04
SP8	5.32	5.01	3.21	4.35	4.89	4.95	6.30	6.58	6.10	6.28	3.60	5.48	3.88	4.18	3.30	3.80	6.06	6.42	6.02	5.44	3.81	4.01	5.86	5.39		2.55
SP9	5.60	5.12	3.00	4.03	4.75	4.43	6.08	6.96	5.35	5.83	3.59	4.58	3.80	3.97	3.56	4.26	5.66	6.00	4.86	5.03	4.47	3.89	5.47	6.04	2.55	
SP10	5.63	4.89	4.35	4.35	5.34	5.09	4.49	6.30	4.97	5.65	5.69	5.39	4.99	4.81	4.57	6.32	5.63	4.26	4.67	4.34	6.22	5.19	5.87	5.84	4.65	4.47

5.5 Caracterización estructural de la región S2 de Esg: Identificación de un α-MoRF

5.5.1 Análisis del ensamble estructural generado durante 54 μs de dinámica molecular

Considerando el promedio estructural del ensamble inicial de la región S2 como referencia, se muestra una conformación α-hélice al comienzo de la simulación (Figura 23A). El porcentaje de estructura secundaria por residuo mostró dos α-hélices, una en el N-terminal entre los residuos 111 al 120, y una segunda en el C-terminal entre los residuos 125 al 150. Durante la simulación el contenido de estructura secundaria mostró un menor porcentaje en comparación al ensamble inicial; la conformación α-hélice del N-terminal fue menor por ~30%, mientras que la conformación α-hélice presente en el C-terminal disminuyó ~10%. Simulaciones de MD previas han reportado que el desplegamiento de una proteína puede ocurrir en una escala de ps[37], y es interesante que durante 54μs de simulación, la conformación del α-hélice es la estructura secundaria más persistente en el C-terminal.

Las estructuras persistentes requieren interacciones que las estabilicen. Para encontrar estas interacciones, se calculó el número de contactos y puentes de hidrógeno a lo largo del ensamble y son presentados en los mapas de calor de la Figura 23B y 23C, respectivamente. La frecuencia de interacción entre pares de residuo es indicada por la barra de color, donde el color azul obscuro representa aquellos contactos con una frecuencia alta, mientras que el color amarillo, representa aquellos contactos de menor frecuencia. Los cuadros de color blanco alejados de la diagonal indican que no se encontraron interacciones entre ese par de residuos. Debido a la flexibilidad de las IDPs, es normal tener varios contactos de largo alcance, los cuales podemos ver en la región amarilla. Los estados transitorios de las IDPs son resultado de la asociación de interacciones de corto y largo alcance entre residuos. Una gran probabilidad de contactos entre residuos cercanos en la secuencia primaria es un sello particular de estructuras helicoidales[42], las cuales se pueden identificar en la región del cuadrado de líneas punteadas (Figura 23B).

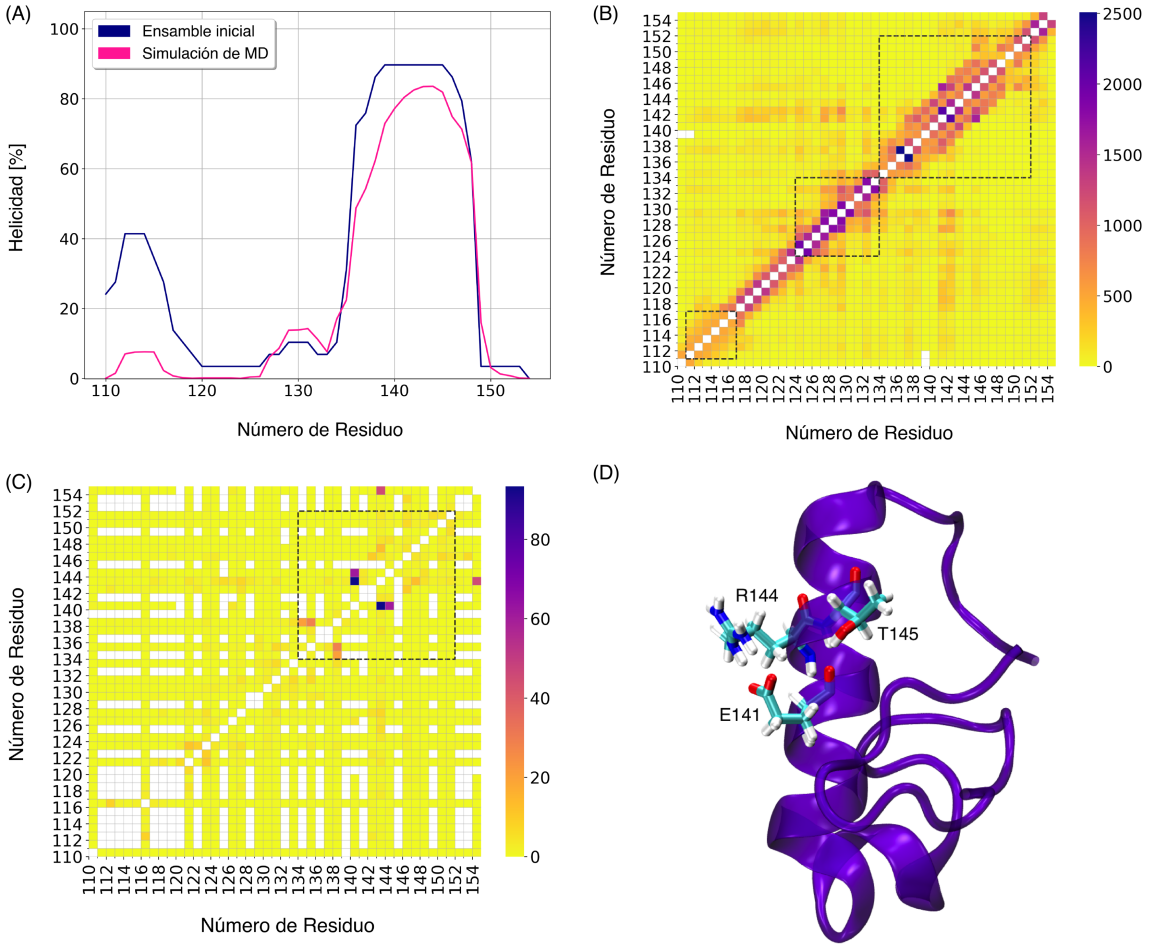


Figura 23. Estructura secundaria y contactos terciarios de la región S2. (A) Porcentaje del tiempo encontrado como hélice para cada residuo. (B) Mapa de calor que representa la interacción de contactos entre pares de residuos. (C) Mapa de calor que representa los puentes de hidrógeno entre pares de residuos. (D) Conformación de α-hélice que representa las interacciones entre los residuos E141 con R144 y T145. La cadena principal se muestra en la representación de listón en color morado, y los aminoácidos son mostrados en varilas en colores CPK.

En las Figuras 23B y 23C, la línea diagonal muestra las interacciones de corto alcance entre residuos. Las conformaciones de α-hélice presentes en el N-terminal de la región S2, entre los residuos 111 a 117, y 124 a 134, no persisten durante la simulación del ensamble como puede observarse en la gráfica de estructura secundaria (Figura 23A). La región con un alto porcentaje de helicidad está presente en el C-terminal de la región S2, entre los residuos 134 a 152, donde ocurre la mayor interacción de contactos entre residuos (Figura 23B) y los puentes de hidrógeno (Figura 23C) de mayor frecuencia. Los puentes de hidrógeno son importantes porque contribuyen a la estabilidad de la estructura

secundaria en proteínas[37]. La conformación de α-hélice pudiera ser estabilizada por interacciones entre los residuos E141 con R144 y T145, los cuales se encuentran a la mitad del α-hélice (Figura 23D) y son las interacciones de puentes de hidrógeno de cadena lateral más frecuentes en el ensamble, mostrados en el cuadro de líneas punteadas (Figuras 23B y 23C). El conjunto de estos resultados indica que la región S2 es una IDR y pudiera ser un α-MoRF que adopta estructura secundaria estable; estos resultados correlacionan con el análisis de predicción del tipo de MoRF (Tabla 8).

5.5.2 Propuesta de mutantes para desestabilizar al α-MoRF

Una de las principales preguntas en este estudio es si la región S2 pudiera ser un α-MoRF. Para responder esta pregunta, se diseñaron mutantes para disminuir la estabilidad del α-hélice. Se encontró que los puentes de hidrógeno más persistentes en el ensamble están entre los residuos E141, R144 y T145. Se ha reportado que los residuos de carga opuesta forman puentes salinos[121,122], estabilizando estructuras locales que pueden ser relevantes en procesos de reconocimiento molecular[42,122–125]. Se sabe que los puentes salinos imparten rigidez local y la disrupción de estas interacciones aumenta la flexibilidad[121,123], los cuales pudieran estar asociados con las transiciones de orden a desorden en IDPs[121]. Las interacciones entre E y R contribuyen favorablemente a la estabilización de hélices[124,125] e interacciones entre E y R (*i+3, i+4*) son más favorables que el puente salino invertido[125]. El Glutamato puede establecer interacciones cadena lateral - cadena principal y cadena lateral – cadena lateral[126]; por su parte, la Treonina tiene una cadena lateral corta y polar, y presenta una fuerte tendencia a formar puentes de hidrógeno con las amidas vecinas de la cadena principal, como se observa en nuestras simulaciones. Los puentes de hidrógeno cadena principal – cadena lateral son determinantes en la formación de hélices y la eliminación de estas interacciones las desestabiliza[126].

Considerando lo anterior, los residuos E141, R144 y T145 fueron sustituidos por residuos alifáticos que promueven el desorden, no forman puentes de hidrógeno y/o tienen una baja propensidad helicoidal, tales como la Valina y Prolina[127]. El residuo E141 fue sustituido por Leucina (E141L), Valina (E141V) y Prolina (E141P). Se generó una doble mutante (E141A_T145A) para eliminar las interacciones entre cadena lateral – cadena principal y cadena lateral – cadena lateral entre E141 y T145. El residuo T145 fue

sustituido por Valina (T145V) y el residuo R144 por Alanina (R144A). Finalmente, Q149P es una variante de *Dme*-Esg presente en la región S2 que no muestra fenotipo[70], fue simulada como control. Por lo tanto, lo que se espera con estas mutantes es poder eliminar las interacciones entre cadena lateral – cadena principal y cadena lateral – cadena lateral, inducir un efecto en la conformación y desestabilizar al α-MoRF.

5.5.3 Análisis de las mutantes

Se seleccionaron cuatro modelos del ensamble inicial (HHP, PI, SP6 y SP9) y fueron generadas las mutantes con CHARMM-GUI a través de la sustitución de la cadena lateral. Cada modelo fue simulado durante 6 μs, para tener un total de 24 μs de cada variante, tanto de la silvestre como de las mutantes. La Figura 24 muestra la estructura secundaria del α-MoRF (residuos 134 al 152) de cada mutante comparada con la silvestre.

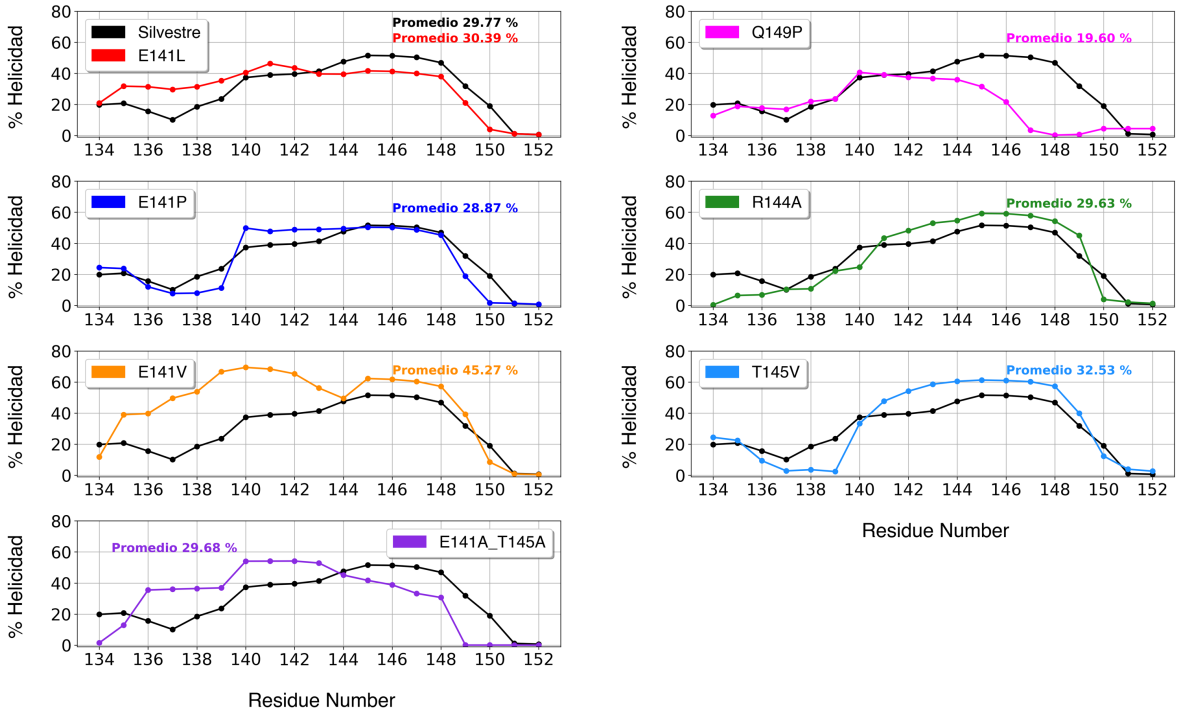


Figura 24. Porcentaje de helicidad por residuo del α-MoRF (residuos 134 a 152) de cada mutante respecto a la silvestre. El valor de helicidad para cada residuo se representa con un punto. El promedio de helicidad para cada variante se indica con su respectivo color.

La mayoría de las mutantes mostraron una diferencia pequeña de helicidad con respecto a la silvestre, ya sea hacia el N-terminal ó C-terminal de la hélice. La doble mutante y Q149P disrumpen el C-terminal del α-MoRF; la pérdida de estructura para Q149P podría indicar que el α-MoRF no necesita extenderse tanto para ser funcional. Romper los

puentes de hidrógeno intramoleculares de un péptido ha sido identificado como un paso clave en el desplegamiento de proteínas[37], pero los puentes de hidrógeno que involucran átomos de cadena lateral en las mutantes no muestran diferencias significativas comparados con la silvestre, y su frecuencia es muy baja. Por otra parte, el número de contactos entre carbono – carbono se incrementó, especialmente entre la región del α-MoRF y el resto de la región de S2; estos contactos no se encuentran presentes de manera significativa en la variante silvestre (Tabla 14). Un problema recurrente con los modelos de solvente implícito es el desbalance de la descripción de términos hidrofóbicos e hidrofílicos. Sin embargo, las mutantes de la región S2 mencionadas anteriormente, parecen estar estabilizadas por interacciones aromáticas. Esto podría deberse a un evento de nucleación causado por el aumento de hidrofobicidad de la hélice a nivel local, ya que todas las mutaciones cambiaron un residuo cargado o polar por uno hidrofóbico.

Las mutantes tienen interacciones en común, involucrando residuos aromáticos presentes en la hélice (Y138, F142 y Y143, rodeados de los residuos mutados) y fuera de ella (Y125, Y128, W130 y F133), los cuales estabilizan la estructura helicoidal (Tabla 14 y Figura 25) a través de la formación de un pequeño núcleo hidrofóbico. La región S2 está enriquecida de residuos aromáticos y prolinas que pudieran estar involucrados también en la estabilización del α-MoRF. Recientemente, se reportó que las interacciones entre prolinas y residuos aromáticos en las posiciones *i±1* e *i-2* pueden formar sitios de nucleación estructural en IDPs[128]. La frecuencia de estas interacciones en las simulaciones se muestra en la Tabla 15, y las estructuras que las representan están incluidas en el Figura 26. Estas interacciones son comunes en las simulaciones y están notablemente enriquecidas en la doble mutante E141A_T145A en todos los pares residuo aromático – prolina, y aproximadamente una vuelta de hélice alejada del sitio de mutación en la mutante E141P. Todos los residuos de prolina en las simulaciones están en configuración *trans*. Para explorar el sesgo de su muestreo conformacional, calculamos el diagrama de Ramachandran para el ensamble silvestre durante los 54 μs de simulación (Figura 27), el cual muestra poblaciones de poliprolina-II y α-hélice, con una pequeña preferencia para esta última.

Tabla 14. Los contactos de largo alcance (n → n + 4 o más) más frecuentes entre la región del α-MoRF (residuos 134 al 152) y el resto de la región S2 de *Dme*-Esg, para cada ensamble durante 24 μs de simulación. La frecuencia fue calculada para todos los pares de átomos de carbono dentro de 6 Å, y luego sumada para todos los pares de carbono - carbono de los residuos que están interaccionando.

Wild-type		E141L		E141P		E141V		E141A-T145A		R144A		T145V		Q149P	
Par	%	Par	%	Par	%	Par	%	Par	%	Par	%	Par	%	Par	%
P123-Y138	94	F133-Y143	676	Y128-Y143	1217	F133-Y143	807	Y128-F142	2148	W130-F142	741	F133-Y143	780	W130-Y138	878
V119-F142	88	W130-F142	598	W130-H134	577	W130-Y138	533	P129-Y138	2042	W130-Y138	726	W130-F142	653	W130-P137	665
P121-Y143	69	V120-F142	548	V119-Y138	473	V120-F142	506	Y128-Y138	1998	Y125-Y138	563	N132-F142	517	W130-F142	560
V118-F142	60	W130-H134	498	Y125-Y138	467	W130-F142	418	Y125-I146	1528	W130-Y143	481	F133-F142	443	W130-R144	553
P121-F142	48	P121-Y138	479	F133-P141	465	W130-Y143	362	Y128-H134	1204	Y128-Y143	417	P123-F142	390	W130-M135	503

Nota: Números que exceden el 100% representan más de un contacto carbono – carbono entre el par de residuos. Por ejemplo, Y128 y F142 presentó un poco más de 21 contactos carbono – carbono en las simulaciones para la doble mutante E141A-T145A.

Tabla 15. Los contactos más frecuentes (i±1 and i-2) entre residuos aromáticos y Prolinas en la región S2 para cada ensamble durante 24 μs de simulación. La frecuencia fue calculada para todos los pares de átomos de carbono dentro de 6 Å, y luego sumada para todos los pares de carbono – carbono de los residuos que están interaccionando.

Sistema	% Frecuencia				
	Y125-P126	Y128-P129	P129-W130	P137-Y138	P141-Y142
Wild-type	1763	1549	1543	2238	-
E141L	1259	1507	1153	1829	-
E141P	1717	1801	763	2572	1895
E141V	1361	1508	1012	1692	-
E141A_T145A	2562	2250	1949	3187	-
Q149P	1576	1640	1047	1758	-
R144A	1462	1232	1165	1788	-
T145V	1432	1237	1248	2193	-

Nota: Números que exceden el 100% representan más de un contacto carbono – carbono entre el par de residuos.

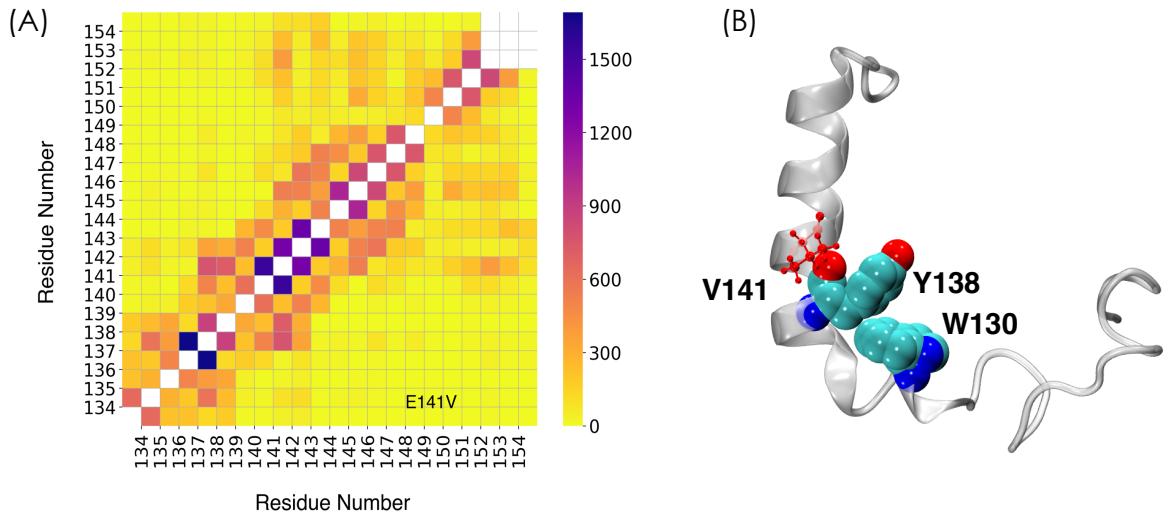


Figura 25. A) Mapa de calor que representa las interacciones residuo-residuo presentes en la región α-MoRF (residuos 134 al 152) de la mutante E141V de la región S2 de *Dme*-Esg. B) Conformación que representa una interacción frecuente de largo alcance de la región α-MoRF con el resto de la región de S2 de Esg. La cadena principal de la proteína se representa en color gris, y los residuos que están interaccionando se representan como esferas de van der Waals en colores CPK. El residuo mutado es mostrado en esferas y varillas de color rojo.

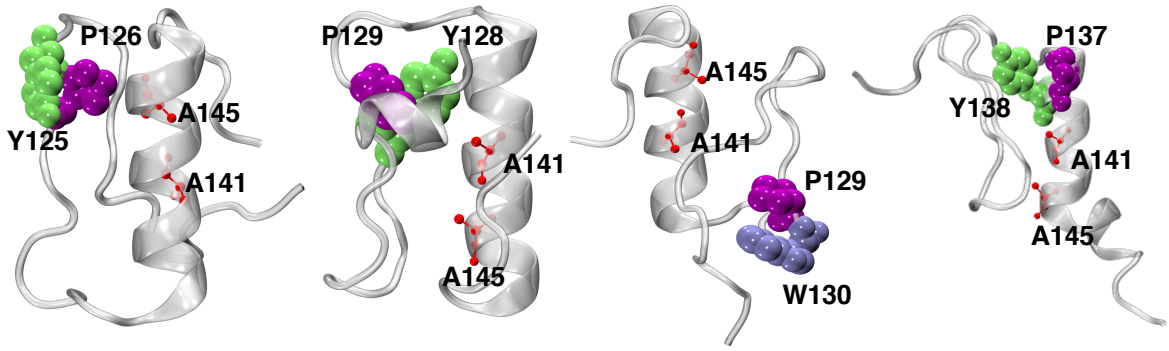


Figura 26. Conformaciones que muestran las interacciones frecuentes entre residuos aromáticos y Prolinas en la región s2 de la mutante E141A_T145A. La cadena principal de la proteína se representa en color gris y los residuos que están interaccionando como esferas de van der Waals: Tirosina en color verde, Prolinas en morado y Triptófano en azul hielo. Los residuos mutados se representan como esferas y varillas de color rojo.

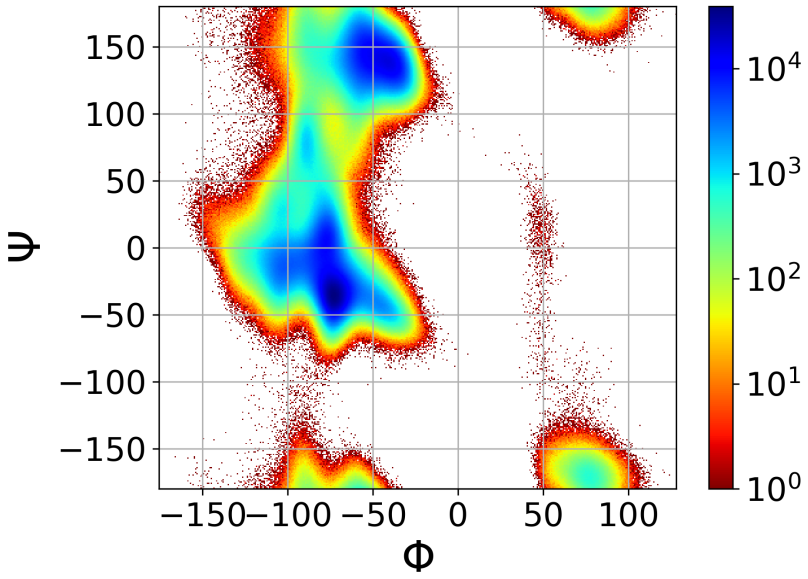


Figura 27. Mapa de Ramachandran para residuos de Prolina (P121, P123, P126, P129, y P137) en el ensamble de S2 de *Dme*-Esg durante 54 *μ*s de simulación.

5.5.4 Propuesta de péptidos para realizar dinámica molecular con CHARMM36m y solvente explícito

Considerando el promedio estructural del ensamble inicial de la región S2 como referencia, y mostrando que las simulaciones realizadas con CHARMM36 y el modelo de solvente GBSA fueron incapaces de desestabilizar fuertemente al α-MoRF, se realizaron simulaciones de la región del α-MoRF (residuos 120 al 152) con el campo de fuerza CHARMM36m y solvente explícito (TIP3P). Es importante mencionar que CHARMM36m ha sido parametrizado para simular IDPs y proteínas ordenadas [99]. Se generó un modelo helicoidal de la variante silvestre utilizando Chimera. Usando CHARMM-GUI se generaron las mutantes E141P y E141L, y de esta manera al mutar Glu por Leu, eliminamos la carga negativa proveniente del E y la interacción del puente salino con R144, y además, aumentamos localmente la hidrofobicidad sin eliminar la posibilidad de seguir estableciendo puentes de hidrógeno en cadena principal. Al mutar Glu por Pro, promovemos el desorden estructural y eliminamos completamente la posibilidad de formar puentes de hidrógeno hacia el extremo N-terminal en cadena principal y hacia el extremo C-terminal utilizando cadena lateral. Las simulaciones se realizaron bajo el esquema de repartición de masas de átomos de hidrógeno con el fin de acelerar el tiempo de muestreo. Se simularon tres trayectorias independientes de cada variante; el tiempo de simulación se muestra en la Tabla 4. Se generó un ensamble de cada variante,

considerando a partir de la primera conformación que presentó 0 % de α-hélice durante al menos 500 ps de cada trayectoria independiente. De esta manera, obtenemos estructuras que no presentan ya el efecto fundador de la estructura inicial (Figura 28). El número de conformaciones presentes en el ensamble de cada variante es distinto. Por ejemplo, la variante silvestre tiene 14,207,000 conformaciones, las cuales provienen de las trayectorias 1 y 2; la mutante E141P presenta 13,374,000 conformaciones, las cuales provienen de las trayectorias 1 y 3; mientras que, la mutante E141L presenta 7,757,000 conformaciones provenientes solo de la trayectoria 2. Debido a que no presentan el mismo número de conformaciones, es necesario simular más tiempo a las mutantes para igualar el número de conformaciones respecto a la silvestre. Sin embargo, considerando estos ensambles se realizó el análisis de fracción helicoidal respecto al tiempo, estructura secundaria por residuo y contactos sobre carbonos.

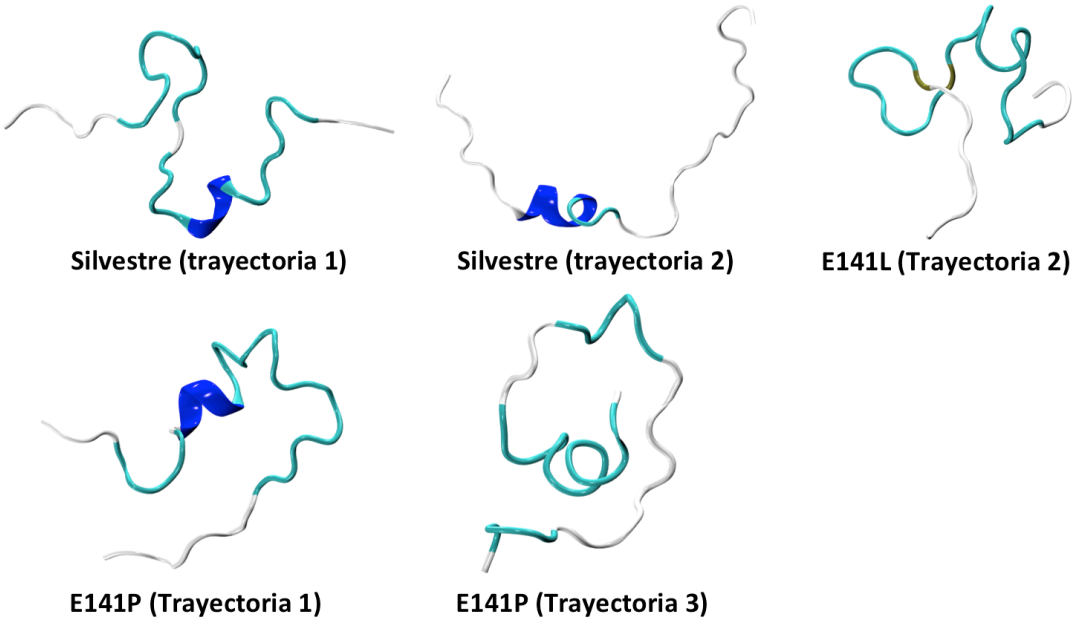


Figura 28. Estructuras representativas del último ps durante los 500 ps que mostraron 0% de α-hélice de cada una de las trayectorias que conforman al ensamble conformacional de la silvestre y mutantes E141L y E141P. Las conformaciones se muestran en la representación de estructura secundaria por default en VMD. En color blanco se muestran las regiones coil, en cian las regiones dobladas (en inglés conocidas como "bend") y en azul las regiones hélices-3_{10}.

5.6 Validación de las simulaciones con CHARMM36m y solvente explícito

5.6.1 Simulación del péptido $(AAQAA)_3$

La simulación de $(AAQAA)_3$ usando CHARMM36m y solvente explícito ya ha sido reportada como se muestra en el Tabla 11, cuyo porcentaje de helicidad es ~17%. Sin embargo, no se ha reportado la simulación de $(AAQAA)_3$ utilizando el esquema de repartición de masas, el cual hemos utilizado en este proyecto. Nuestra simulación de $(AAQAA)_3$ comenzó de una conformación 100% helicoidal, y fue simulada durante 16 μs usando el mismo protocolo de simulación para los modelos del α-MoRF de *Dme*-Esg. La Figura 29 muestra la fracción de α-hélice calculada cada 10 ns, en donde se observa que $(AAQAA)_3$ intercambia frecuentemente conformaciones coil y helicoidales a partir del primer microsegundo de simulación. Así mismo, se calculó el porcentaje de helicidad por residuo (Figura 29), el cual presenta un promedio de ~23.84 %, ligeramente alto a lo reportado en condiciones experimentales, pero dentro del margen esperado de helicidad (Tabla 11).

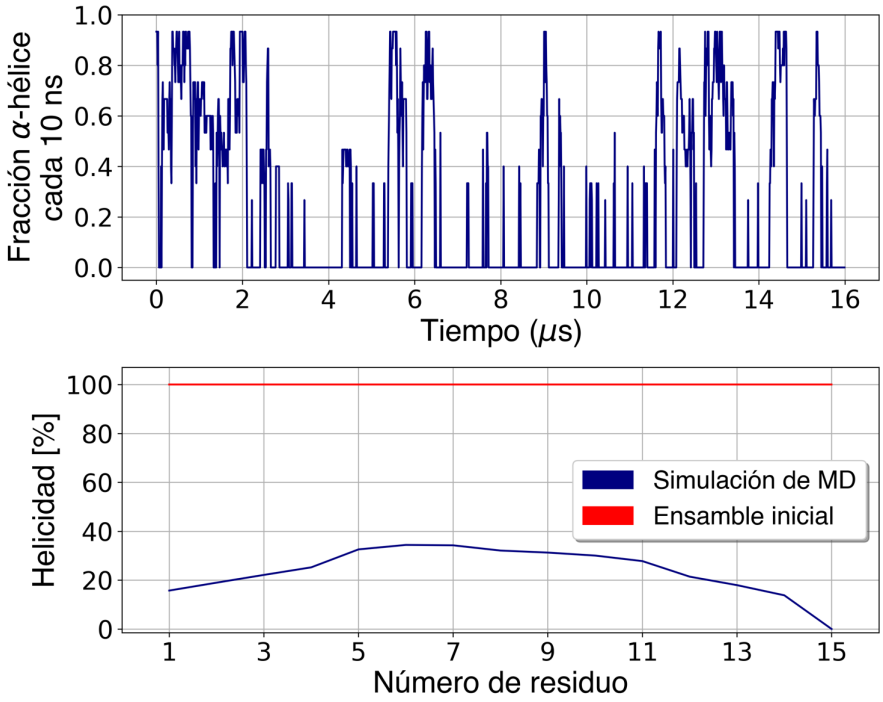


Figura 29. Ensamble de $(AAQAA)_3$ durante 16 μs of MD simulación. (A) Fracción de α-hélice de $(AAQAA)_3$ calculado en bloques de 10ns. (B) Porcentaje de helicidad por residuo, promedio sobre los 16 μs (línea azul) comparada con el ensamble inicial (línea roja).

5.7 Caracterización estructural del efecto de las mutantes presentes en el α-MoRF con CHARMM36m y solvente explícito.

5.7.1 Análisis de estructura secundaria

El resultado de la fracción de α-hélice de cada variante se muestra en la Figura 30. En cada uno de los ensambles se puede observar que la fracción de helicidad para la silvestre y las mutantes no sobrepasa el 40% de helicidad. Además, en la silvestre y en las mutantes se puede observar la presencia de estructuras que presentan una ruptura total del α-MoRF, siendo la mutante E141P la que presenta con mayor frecuencia estructuras que presentan 0% de helicidad. Se calculó el porcentaje de la fracción de helicidad para cada variante y se compararon los resultados con los resultados de dicroísmo circular (CD) obtenidos por Ángel Peláez y la Dra. Lina Rivillas, al determinar estructura secundaria utilizando Trifluoretano (TFE). TFE ha sido utilizado en el estudio de péptidos por su capacidad de estabilizar elementos de estructura secundaria. TFE puede inducir y promover la formación de hélices, principalmente si la secuencia del péptido está sesgada a adoptar conformaciones helicoidales[129,130]. Utilizando porcentajes pequeños de TFE no se pudo obtener el contenido de α-hélice suficiente para ser cuantificable, por lo que se optó utilizar 100% de TFE para obtener un mínimo de α-hélice evaluable, con la certeza que solo la región de la secuencia que es capaz de adoptar estructura helicoidal se vería afectada por el inductor de hélice. Los resultados de CD al utilizar 100% de TFE se muestran en la Tabla 16.

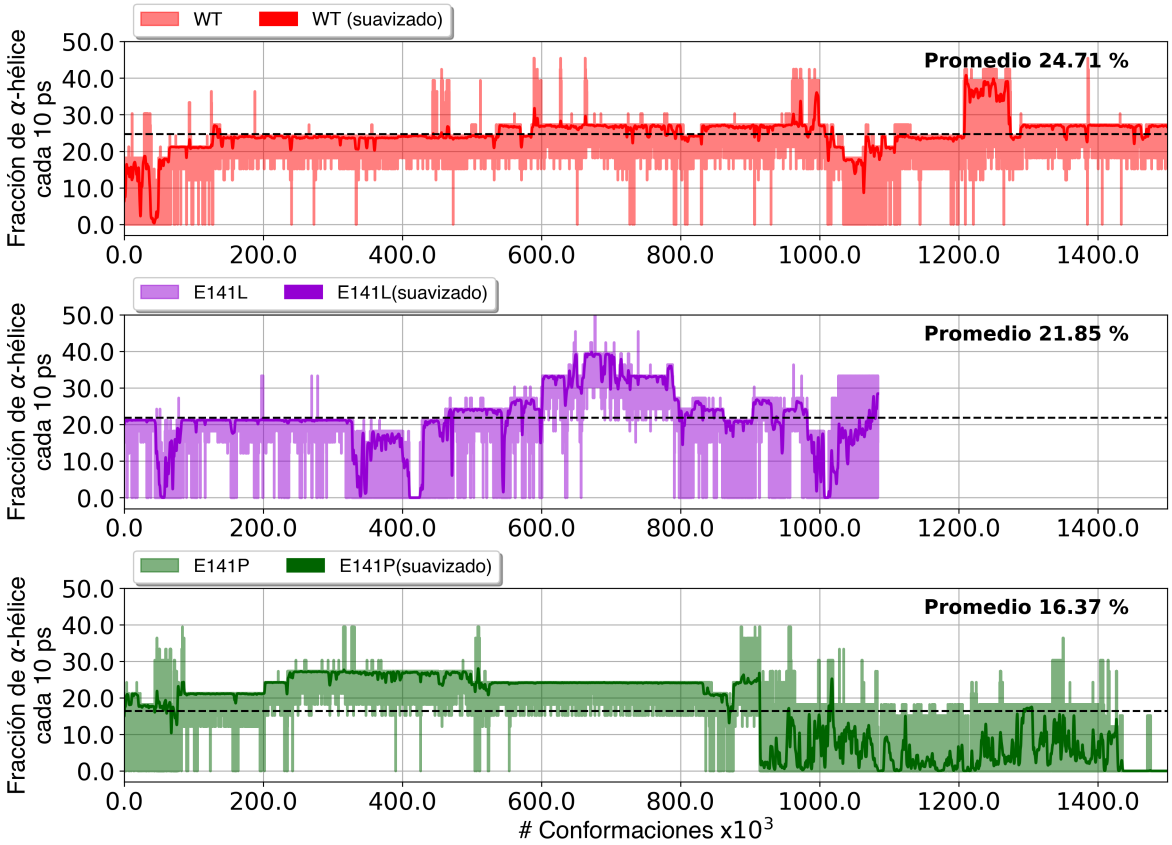


Figura 30. Fracción de α-hélice graficada cada 10 ps del ensamble obtenido con CHARMM36m y solvente explícito de la variante Silvestre y las mutantes E141P y E141L. La línea punteada indica el promedio de α-hélice en cada condición.

Los resultados de las simulaciones son cualitativamente consistentes con los datos experimentales. Sin embargo, el porcentaje de helicidad para las mutantes aún están por encima del porcentaje de helicidad obtenido con CD. Esperamos que mejorando el muestreo estas tendencias se acerquen más a la medida experimental.

Tabla 16. Porcentaje de la fracción de helicidad obtenido de los ensambles de las variantes silvestre y mutantes E141L y E141P.

Variante	% de α-hélice obtenida con simulación de MD	% de α-hélice obtenida con CD usando 100 % de TFE
silvestre	25	22
E141L	23	16
E141P	16	8

Además, se calculó el porcentaje de estructura secundaria por residuo. En la Figura 31 se muestra el porcentaje de α-hélice por residuo del ensamble obtenido para cada variante. En el eje de las abscisas se muestra la secuencia de la región S2 en su variante silvestre, resaltando en negrita el residuo 141 que corresponde a un Glu y el cual ha sido mutado por Leu en el caso de E141L y por Pro en el caso de E141P.

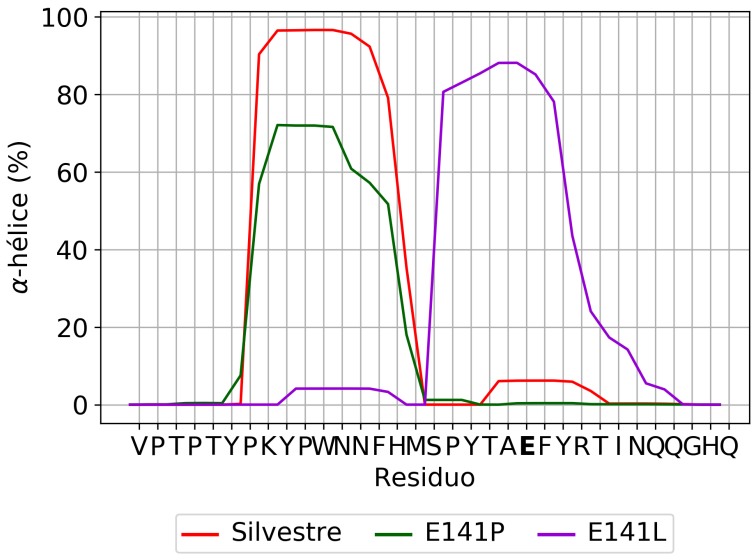


Figura 31. Porcentaje de α-hélice por residuo del ensamble obtenido de la variante silvestre y las mutantes E141P y E141L.

Los resultados muestran que la variante silvestre pierde helicidad hacia los extremos N-terminal y C-terminal del α-MoRF, el cual disminuye a 0%, mientras que persiste un α-hélice de hasta 100% de helicidad entre los residuos 125 al 135. Por su parte, la mutante E141P disminuyó la helicidad hacia sus extremos N-terminal y C-terminal a 0%; sin embargo mantuvo poco más de 70% de helicidad en la misma región que mantiene helicidad la silvestre (residuos 125 al 135). Así mismo, la mutante E141L disminuyó la helicidad de sus extremos N-terminal y C-terminal; sin embargo mantiene helicidad poco más del 80% hacia el C-terminal, entre los residuos 135 al 144. La región donde persiste la estructura de α-hélice es una región rica en residuos de Pro y residuos aromáticos, los cuales pudieran estar estableciendo interacciones para estabilizar el α-MoRF.

5.7.2 Contactos terciarios

Para conocer la estabilidad del α-MoRF se calculó el mapa de contactos para cada variante, los cuales representan las interacciones entre átomos de carbonos cuya distancia es menor o igual a 6 Å (Figura 33). El cuadrado de línea punteada representa la región donde se observa la α-hélice, la cual se localiza entre los residuos 125 al 135 para la silvestre y E141P, y entre los residuos 135 al 144 para E141L (Figura 31 y Figura 33). Así mismo, se pueden identificar interacciones de largo alcance que muestran la interacción entre los dominios N-terminal y C-terminal, las cuales ocurren en las tres variantes (Figura 33). A su vez, las interacciones más frecuentes en la variante silvestre y en las mutantes E141L y E141P son entre aminoácidos hidrofóbicos y polares, los cuales contribuyen a la estabilización del α-MoRF (Figura 32) hacia el N-terminal para la variante silvestre y mutante E141P, y hacia el C-terminal para la mutante E141L (Tabla 17).

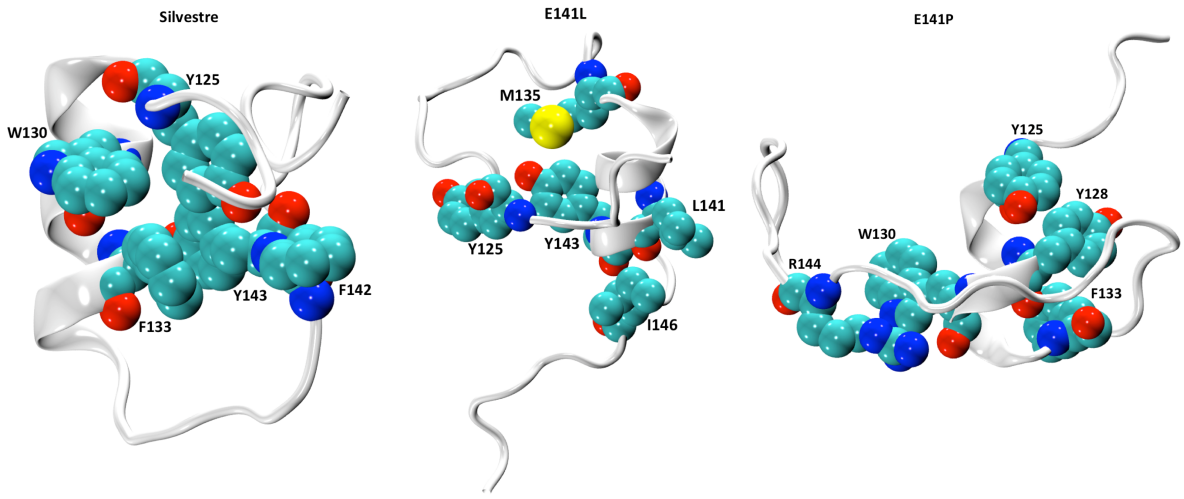


Figura 32. Estructuras representativas del α-MoRF presente en la silvestre, E141L y E141P. La cadena principal se muestra en la representación listón en color blanco, y los aminoácidos hidrofóbicos y polares son mostrados como esferas de van der Waals en colores CPK, los cuales establecen las interacciones de mayor frecuencia que estabilizan al α-MoRF.

La región del α-MoRF es rica en residuos de Pro y residuos aromáticos; estos últimos son representados con el símbolo '★' en el Figura 33. Como se mencionó anteriormente, se ha reportado que las interacciones entre residuos de Pro y residuos aromáticos en las posiciones *i±1* e *i-2* pueden formar sitios de nucleación estructural en IDPs[128]. La

frecuencia de estas interacciones en las simulaciones se puede observar que son comunes en la silvestre, E141L y E141P (Figura 33 y Tabla 18), y en los tres sistemas están igualmente enriquecidas sin mostrar diferencias significativas. Así mismo, se evaluaron las interacciones de mayor frecuencia que establece el residuo 141 (Figura 34 y Tabla 19). Se puede observar que el residuo 141 en E141L establece interacciones de mayor frecuencia, seguida de la silvestre y E141P. Además, las interacciones que establece el residuo 141 en E141L son en su mayoría con residuos aromáticos y mantiene la interacción con R144, el cual se presenta con menor frecuencia en la silvestre y en la mutante E141P.

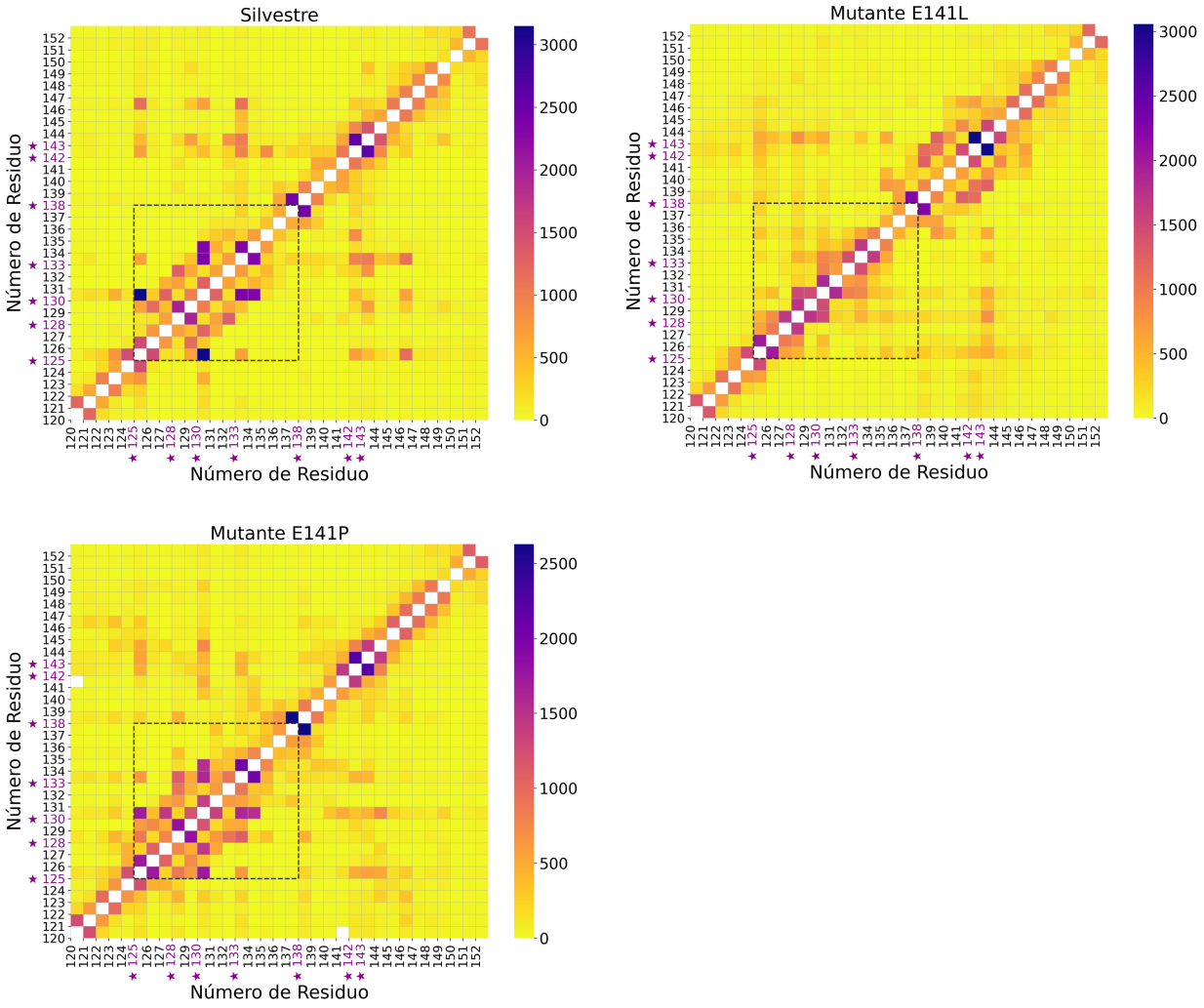


Figura 33. Mapa de calor que representa las interacciones entre átomos de carbonos dentro de una distancia de 6Å del ensamble obtenido con CHARMM36m y solvente explícito de la variante Silvestre y las mutantes E141P y E141L. La línea diagonal muestra las interacciones de corto alcance. En los cuadros de líneas punteadas se representan las interacciones de cadena principal entre residuos que conservan helicidad.

Tabla 17. Los contactos de largo alcance (n → n + 5 o más) más frecuentes del α-MoRF (residuos 120 al 152) de la región S2 de *Dme*-Esg presentes en la variante Silvestre y mutantes E141P y E141L, obtenidos de la simulación de MD con CHARMM36m y solvente explícito. La frecuencia fue calculada para todos los pares de átomos de carbono dentro de 6 Å, y luego sumada para todos los pares de carbono - carbono de los residuos que están interaccionando.

WT		E141L		E141P	
Residuos	%	Residuos	%	Residuos	%
Y125-W130	3149	Y128-Y143	797	Y125-W130	1724
Y125-I146	1167	M135-Y143	583	Y128-F133	1104
F133-I146	1124	Y125-Y143	573	W130-R144	753
F133-Y143	1022	W11-Y143	564	Y125-F133	624
F133-F142	955	L141-I146	535	Y125-Y143	557

Nota: Números que exceden el 100% representan más de un contacto carbono – carbono entre el par de residuos.

Tabla 18. Los contactos más frecuentes (i±1 y i-2) entre residuos aromáticos y Prolinas en la región del α-MoRF presentes en la variante Silvestre y mutantes E141P y E141L. La frecuencia fue calculada para todos los pares de átomos de carbono dentro de 6 Å, y luego sumada para todos los pares de carbono – carbono de los residuos que están interaccionando.

Sistema	% Frecuencia				
	Y125-P126	Y128-P129	P129-W130	P137-Y138	P141-Y142
silvestre	1526	1971	987	2446	-
E141L	2000	1731	1479	2255	-
E141P	1745	1808	1198	2628	1430

Tabla 19. Los contactos más frecuentes que establece el residuo 141 en la variante Silvestre, y las mutantes E141P y E141L, obtenidos de la simulación de MD con CHARMM36m y solvente explícito. La frecuencia fue calculada para todos los pares de átomos de carbono dentro de 6 Å, y luego sumada para todos los pares de carbono - carbono de los residuos que están interaccionando.

WT		E141L		E141P	
Residuos	%	Residuos	%	Residuos	%
E141-F142	1059	L141-F143	1450	P141-F143	1430
A140-E141	688	Y138-L141	1246	A140-P141	599
E141-Y143	638	L141-R144	853	W130-P141	488
T139-E141	628	A140-L141	677	P141-Y143	472
E141-R144	276	L141-I146	555	P141-R144	247

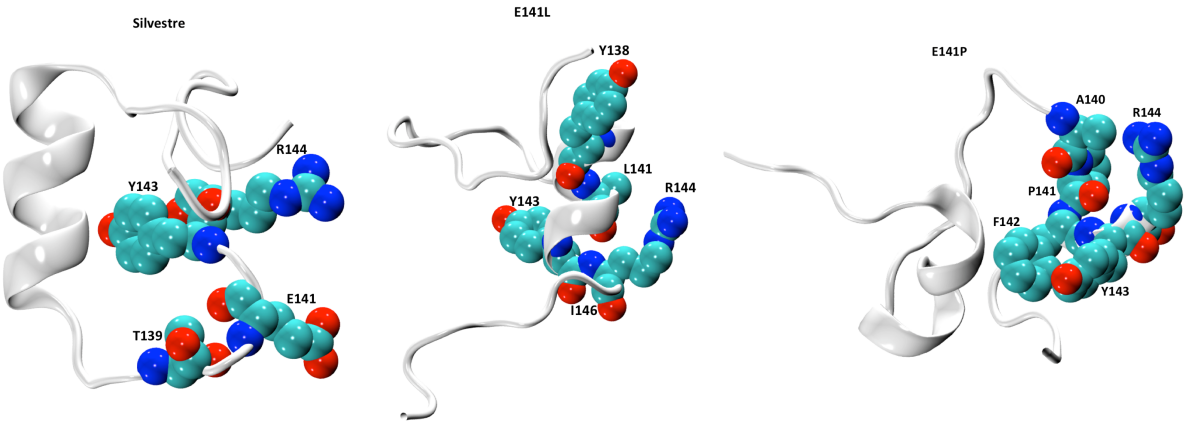


Figura 34. Estructuras representativas de las interacciones que establece el residuo 141 presente en la silvestre, E141L y E141P. La cadena principal se muestra en la representación listón en color blanco, y los aminoácidos hidrofóbicos y polares son mostrados como esferas de van der Waals en colores CPK, los cuales establecen las interacciones de mayor frecuencia con el residuo 141.

El conjunto de estos resultados sugiere que las mutantes E141P y E141L son capaces de romper el α-MoRF en ambos extremos. Es importante mencionar que el tiempo de simulación generado para cada variante no garantiza un muestreo suficiente, ya que deberíamos observar una convergencia estructural en relación con el extremo N-terminal en las tres variantes, debido a que son idénticos en secuencia en su extremo N-terminal y se espera que tengan el mismo comportamiento estructural. Además, es una región rica en residuos de Pro, los cuales son promotores de desorden. Lo anterior probablemente se deba a que las estructuras quedaron atrapadas en un mínimo local, siendo incapaces de brincar la barrera energética que lo separa de otros subestados conformacionales. Estos resultados a su vez reflejan la importancia de generar varias simulaciones cortas iniciando de diferentes modelos estructurales en vez de una simulación larga con un solo modelo para explorar el ensamble conformacional de una IDP y/o IDR, debido a que esta estrategia permite un muestreo conformacional más eficiente e incrementa la probabilidad para converger con datos experimentales.

6.- Conclusiones

Esg es un factor de transcripción de la familia Snail expresado en *Drosophila melanogaster (Dme)* y está involucrado en diferentes funciones, entre ellas el desarrollo del sistema nervioso. Estructuralmente, el C-terminal de la familia Snail es una región conservada que tiene ZNFs e interactúa con ácidos nucleicos, mientras que el N-terminal es divergente y actualmente, hay poca información funcional y estructural; sin embargo, se le han asociado funciones como degradación de proteínas, donde los ZNFs no son necesarios. De acuerdo a los análisis realizados y resultados obtenidos en este proyecto concluimos lo siguiente:

- Hemos comparado la secuencia de Esg de *Dme* y proteínas homólogas para conocer el grado de conservación. Los resultados sugieren que tanto su extremo N-terminal como su C-terminal se encuentran conservados de distintas maneras. Al comparar las secuencias de Esg de *Dme* con proteínas ortólogas que pertencen a diferentes filos, los resultados sugieren que a nivel de secuencia el N-terminal no está conservado pero a nivel de desorden sí, con lo cual pudiera definirse el N-terminal de Esg como una IDR de desorden flexible. En el perfil de desorden hemos identificado regiones propensas a ordenarse presentes en el N-terminal de la mayoría de las proteínas analizadas. Estas IDRs presentan propiedades estructurales de una proteína ordenada, y las cuales pudieran adoptar conformaciones de α-hélice y β-plegada. Sin embargo, se sugiere ampliar el análisis estructural de estas regiones IDRs para confirmar si pudieran ser α-MoRFs o β-MoRFs.

- Las simulaciones computacionales han demostrado obtener una buena descripción del ensamble conformacional de IDPs, incluso cuando la mayoría de los campos de fuerza han sido desarrollados para proteínas plegadas. En este trabajo, se investigó la presencia de MoRFs en el N-terminal desordenado del factor de transcripción Esg de *Dme* por simulaciones de MD usando CHARMM36 y el modelo de solvente GBSA. Hemos simulado la región S2 (residuos 111 a 155) presente en el N-terminal de *Dme*-Esg. El análisis del ensamble conformacional de esta región mostró la presencia de estructura residual que pudiera ser un α-MoRF, la cual se estabiliza por contactos terciarios y persiste durante 54 μs de simulación de MD.

- Mutaciones puntuales en la región S2 fueron construidas para desestabilizar el probable α-MoRF, probando su estabilidad durante 24 μs de cada una. Los resultados de las mutantes no mostraron una diferencia significativa de helicidad del α-MoRF con respecto a la silvestre, a excepción de la mutante Q149P que mostró pérdida de estructura residual hacia el C-terminal y lo cual podría indicar que el α-MoRF no necesita extenderse tanto para ser funcional.

- Como un control de las simulaciones, se evaluó la población de α_L durante los 54 μs de simulación de MD, el cual fue de 1.36% y se encontró que es significativamente más bajo al reportado en ensambles de IDPs generados con CHARMM36. Así mismo, se realizó una matriz de RMSD 2D para evaluar el tiempo necesario de simulación para observar convergencia estructural entre las trayectorias de los modelos simulados de la región S2. Los resultados mostraron que 2 μs de simulación es tiempo suficiente para observar convergencia estructural entre modelos, y de esta manera, asumimos que las estructuras generadas por los 54 μs de simulación de MD pueden representar, como una buena primera aproximación, la diversidad del espacio conformacional de la región S2 de *Dme*-Esg.

- Se reportaron 32 μs de simulación de $(AAQAA)_3$ y su promedio de contenido helicoidal durante el ensamble fue de ~24.53% (un valor cercano a datos experimentales), el cual indica que CHARMM36 y GBSA proporcionan un buen balance entre las conformaciones hélice y coil para este sistema, pero la combinación de estas herramientas computacionales fue insuficiente para mostrar el efecto de mutaciones puntuales que deberían haber desestabilizado el α-MoRF. El conjunto de resultados sugiere que el campo de fuerza de CHARMM36 y el modelo de solvente GBSA pueden generar ensambles estructurales con expresión y estabilización de estructuras helicoidales gracias a interacciones terciarias, pero, la descripción de estructura secundaria y el balance entre interacciones hidrofóbicas e hidrofílicas son satisfactorios.

- Por otra parte, se simuló la variante silvestre y las mutantes E141P y E141L de la región del α-MoRF (residuos 120 a 152) presente en la región S2 de *Dme*-Esg a partir de modelos helicoidales. Las simulaciones de MD se realizaron con CHARMM36m y solvente explícito (TIP3P), bajo el esquema de repartición de masas de átomos de

hidrógeno, con el fin de acelerar el tiempo de muestreo. Se realizaron tres trayectorias independientes de cada variante, y se obtuvo un ensamble considerando a partir de la primera conformación que presentó 0 % de α-hélice al menos 500 ps en cada una de las trayectorias. Los resultados mostraron que la mutantes E141P y E141L son capaces de desestabilizar al α-MoRF en ambos extremos (N-terminal y C-terminal). Sin embargo, en las mutantes aún se puede observar la presencia de α-hélice en una región rica en residuos de Pro y residuos aromáticos, los cuales establecen interacciones para estabilizar al α-MoRF. Así mismo, se observó que las interacciones más frecuentes se establecen entre aminoácidos hidrofóbicos y polares, las cuales también contribuyen a la estabilización del α-MoRF. Sin embargo, los resultados indican que el tiempo de muestreo es aún insuficiente al no observar convergencia estructural en los ensambles de la variante silvestre y las mutantes, debido a que son idénticos en secuencia en su extremo N-terminal y se espera que tengan el mismo comportamiento estructural, además de ser una región rica en residuos de Pro, los cuales son promotores de desorden. Lo anterior probablemente se deba a que las estructuras quedaron atrapadas en un mínimo local.

- Respecto a las metologías utilizadas en este proyecto:

  - Utilizar solvente implícito resulta una manera útil para simular tiempos largos al disminuir el costo computacional, por el hecho de solo considerar de manera explícita los átomos del soluto. Esto permite que la búsqueda conformacional sea más rápida (a diferencia del solvente explícito), y haya una menor precisión en los resultados al no considerar las interacciones entre el soluto y solvente, las cuales pudieran ser importantes en el proceso de plegamiento. Además, se ha reportado que utilizar solvente implícito se generan conformaciones más estructuradas y compactas.

  - Por su parte, utilizar solvente explícito permite obtener simulaciones con mayor precisión (dependiendo del modelo de agua utilizado) y correlación con datos experimentales; sin embargo no resultan ser la primera opción para simular tiempos largos de sistemas de IDPs y/o IDRs ya que el costo computacional es alto. A su vez, utilizar solvente explícito con el método de repatición de masas, permite aumentar el tiempo de integración y reducir el costo computacional.

- Finalmente, simular IDPs y/o IDRs no es una tarea fácil. De acuerdo a los reportes de las simulaciones realizadas de IDPs a la fecha y al conjunto de resultados obtenidos en este proyecto, se refleja la importancia de generar varias simulaciones cortas iniciando de diferentes modelos estructurales, en vez de una simulación larga con un solo modelo para explorar el ensamble conformacional de una IDP y/o IDR, debido a que esta estrategia permite un muestreo conformacional más eficiente e incrementa la probabilidad para converger con datos experimentales. Actualmente no existe un protocolo específico para simular IDPs, pero existe una continua búsqueda para desarrollar y mejorar los campos de fuerza y puedan ser aplicados tanto a proteínas plegadas como desordenadas.

7.- Perspectivas

1.- Realizar un análisis estructural de las regiones propensas a ordenarse presentes en el N-terminal del factor de transcripción de Esg de Dme y sus ortólogos, con el fin de determinar si son MoRFs con probabilidad a sufrir modificaciones postraduccionales, y describir el papel biológico que pudieran tener.

2.- Seleccionar cuatro modelos estructurales de la región S2 de *Dme*-Esg obtenidos de las simulaciones de las mutaciones, las cuales fueron simuladas con CHARMM36 y modelo del solvente GBSA. Se requiere que las estructuras presenten una RMSD mínima de 10 Å para asegurar su diversidad estructural, y arrancar nuevas simulaciones a partir de ellas.

3.- Generar simulaciones con CHARMM36m y solvente explícito (TIP3P) de la región del α-MoRF, utilizando la estrategia de repartición de masas de átomos de hidrógeno y las estructuras obtenidas a partir de la descripción del punto anterior, con el fin de mejorar el muestreo conformacional y caracterizar el ensamble del α-MoRF.

4.- Aprender a generar simulaciones con CHARMM36m y TFE de la región del α-MoRF, utilizando la estrategia de repartición de masas de átomos de hidrógeno y las estructuras obtenidas a partir de la descripción del punto dos, con el fin de igualar las condiciones experimentales y caracterizar el ensamble del α-MoRF.

5.-Actualmente, existe una colaboración directa con los laboratorios del Dr. Carlos Amero y la Dra. Lina Rivillas. En el laboratorio del Dr. Carlos Amero se realizan experimentos de

RMN para la identificación de estructura residual del α-MoRF y por DLS para establecer radio hidrodinámico. En el laboratorio de la Dra. Lina Rivillas se llevó a cabo la síntesis de la variante silvestre y las mutantes E141P y E141L (identificadas a través de las simulaciones), para caracterizar su estructura secundaria por CD. Lo que resta por hacer es mejorar el muestreo conformacional, para establecer la mejor correspondencia posible con los datos experimentales.

8.- Bibliografía

(1) Mannige R V. Dynamic new world: Refining our view of protein structure, function and evolution. Proteomes. **2014**; 2: 128–153.
(2) Van Der Lee R, Buljan M, Lang B *et al.* Classification of intrinsically disordered regions and proteins. Chem. Rev. **2014**; 114: 6589–6631.
(3) Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. Nat. Rev. Mol. Cell Biol. **2015**; 16: 18–29.
(4) Habchi J, Tompa P, Longhi S, Uversky VN. Introducing protein intrinsic disorder. Chem. Rev. **2014**; 114: 6561–6588.
(5) Oldfield CJ, Uversky VN, Dunker AK, Kurgan L. Introduction to intrinsically disordered proteins and regions. In: *Intrinsically Disordered Proteins*. Elsevier, 2019, pp 1–34.
(6) Uversky VN. Natively unfolded proteins: A point where biology waits for physics. *Protein Sci* **2002**; 11: 739–756.
(7) Burger VM, Gurry T, Stultz CM. Intrinsically disordered proteins: Where computation meets experiment. Polymers (Basel). **2014**; 6: 2684–2719.
(8) Das P, Matysiak S, Mittal J. Looking at the Disordered Proteins through the Computational Microscope. *ACS Cent Sci* **2018**; 4: 534–542.
(9) Uversky VN. The multifaceted roles of intrinsic disorder in protein complexes. FEBS Lett. **2015**; 589: 2498–2506.
(10) Uversky VN. Intrinsically disordered proteins from A to Z. Int. J. Biochem. Cell Biol. **2011**; 43: 1090–1103.
(11) Yan J, Dunker AK, Uversky VN, Kurgan L. Molecular recognition features (MoRFs) in three domains of life. *Mol Biosyst* **2016**; 12: 697–710.
(12) Perez A, Morrone JA, Dill KA. Accelerating physical simulations of proteins by leveraging external knowledge. Wiley Interdiscip. Rev. Comput. Mol. Sci. **2017**; 7. doi:10.1002/wcms.1309.
(13) Gianni S, Dogan J, Jemth P. Distinguishing induced fit from conformational selection. *Biophys Chem* **2014**; 189: 33–39.
(14) Keppel TR, Weis DD. Mapping residual structure in intrinsically disordered proteins at residue resolution using millisecond hydrogen/deuterium exchange and residue averaging. *J Am Soc Mass Spectrom* **2015**; 26: 547–554.
(15) Toto A, Malagrinò F, Visconti L *et al.* Templated folding of intrinsically disordered proteins. J. Biol. Chem. **2020**; 295: 6586–6593.
(16) Das RK, Ruff KM, Pappu R V. Relating sequence encoded information to form and function of intrinsically disordered proteins. Curr. Opin. Struct. Biol. **2015**; 32: 102–112.
(17) Jensen MR, Zweckstetter M, Huang JR, Blackledge M. Exploring free-energy

landscapes of intrinsically disordered proteins at atomic resolution using NMR spectroscopy. Chem. Rev. **2014**; 114: 6632–6660.
(18) Uversky VN. Under-folded proteins: Conformational ensembles and their roles in protein folding, function, and pathogenesis. Biopolymers. **2013**; 99: 870–887.
(19) DeForte S, Uversky VN. Order, disorder, and everything in between. Molecules. **2016**; 21. doi:10.3390/molecules21081090.
(20) Schor M, Mey ASJS, MacPhee CE. Analytical methods for structural ensembles and dynamics of intrinsically disordered proteins. Biophys. Rev. **2016**; 8: 429–439.
(21) Ouyang Y, Zhao L, Zhang Z. Characterization of the structural ensembles of p53 TAD2 by molecular dynamics simulations with different force fields. *Phys Chem Chem Phys* **2018**; 20: 8676–8684.
(22) Fadda E, Nixon MG. The transient manifold structure of the p53 extreme C-terminal domain: Insight into disorder, recognition, and binding promiscuity by molecular dynamics simulations. *Phys Chem Chem Phys* **2017**; 19: 21287–21296.
(23) Mauri F, McNamee LM, Lunardi A *et al.* Modification of Drosophila p53 by SUMO modulates its transactivation and pro-apoptotic functions. *J Biol Chem* **2008**; 283: 20848–20856.
(24) Holehouse AS, Sukenik S. Controlling Structural Bias in Intrinsically Disordered Proteins Using Solution Space Scanning. *J Chem Theory Comput* **2020**; 16: 1794–1805.
(25) Fuxreiter M, Simon I, Friedrich P, Tompa P. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J Mol Biol* **2004**; 338: 1015–1026.
(26) Gruet A, Dosnon M, Blocquel D *et al.* Fuzzy regions in an intrinsically disordered protein impair protein-protein interactions. *FEBS J* **2016**; 283: 576–594.
(27) Zhou J, Oldfield CJ, Huang F *et al.* Identification of intrinsic disorder in complexes from Protein Data Bank. *ACS Omega* **2018**; : 1–1.
(28) Shabane PS, Izadi S, Onufriev A V. General Purpose Water Model Can Improve Atomistic Simulations of Intrinsically Disordered Proteins. *J Chem Theory Comput* **2019**; 15: 2620–2634.
(29) Kukharenko O, Sawade K, Steuer J, Peter C. Using Dimensionality Reduction to Systematically Expand Conformational Sampling of Intrinsically Disordered Peptides. *J Chem Theory Comput* **2016**; 12: 4726–4734.
(30) Best RB. Emerging consensus on the collapse of unfolded and intrinsically disordered proteins in water. Curr. Opin. Struct. Biol. **2020**; 60: 27–38.
(31) Chong S-H, Chatterjee P, Ham S. Computer Simulations of Intrinsically Disordered Proteins. *Annu Rev Phys Chem* **2017**; 68: 117–34.
(32) Bottaro S, Lindorff-Larsen K, Best RB. Variational optimization of an all-atom implicit solvent force field to match explicit solvent simulation data. *J Chem Theory Comput* **2013**; 9: 5641–5652.
(33) Onufriev A. Chapter 7 Implicit Solvent Models in Molecular Dynamics Simulations: A Brief Overview. In: *Annual Reports in Computational Chemistry*. Elsevier BV, 2008, pp 125–137.
(34) Zhang W, Ganguly D, Chen J. Residual structures, conformational fluctuations, and electrostatic interactions in the synergistic folding of two intrinsically disordered proteins. *PLoS Comput Biol* **2012**; 8: e1002353.
(35) Ilizaliturri-Flores I, Correa-Basurto J, Bello M *et al.* Mapping the intrinsically disordered properties of the flexible loop domain of Bcl-2: a molecular dynamics simulation study. *J Mol Model* **2016**; 22: 98.

(36) Cino EA, Choy WY, Karttunen M. Characterization of the Free State Ensemble of the CoRNR Box Motif by Molecular Dynamics Simulations. *J Phys Chem B* **2016**; 120: 1060–1068.
(37) Navarro-Retamal C, Bremer A, Alzate-Morales J *et al.* Molecular dynamics simulations and CD spectroscopy reveal hydration-induced unfolding of the intrinsically disordered LEA proteins COR15A and COR15B from: Arabidopsis thaliana. *Phys Chem Chem Phys* **2016**; 18: 25806–25816.
(38) Gong H, Zhang S, Wang J, Gong H, Zeng J. Constructing structure ensembles of intrinsically disordered proteins from chemical shift data. In: *Journal of Computational Biology*. Mary Ann Liebert Inc., 2016, pp 300–310.
(39) Guo X, Han J, Luo R, Chen HF. Conformation dynamics of the intrinsically disordered protein c-Myb with the ff99IDPs force field. *RSC Adv* **2017**; 7: 29713–29721.
(40) Siwy CM, Lockhart C, Klimov DK. Is the Conformational Ensemble of Alzheimer's Aβ10-40 Peptide Force Field Dependent? *PLoS Comput Biol* **2017**; 13: e1005314.
(41) Han M, Xu J, Ren Y. Sampling conformational space of intrinsically disordered proteins in explicit solvent: Comparison between well-tempered ensemble approach and solute tempering method. *J Mol Graph Model* **2017**; 72: 136–147.
(42) Carballo-Pacheco M, Strodel B. Comparison of force fields for Alzheimer's A β42: A case study for intrinsically disordered proteins. *Protein Sci* **2017**; 26: 174–185.
(43) Sherry KP, Das RK, Pappu R V., Barrick D. Control of transcriptional activity by design of charge patterning in the intrinsically disordered RAM region of the Notch receptor. *Proc Natl Acad Sci U S A* **2017**; 114: E9243–E9252.
(44) Harris J, Shadrina M, Oliver C, Vogel J, Mittermaier A. Concerted millisecond timescale dynamics in the intrinsically disordered carboxyl terminus of γ-tubulin induced by mutation of a conserved tyrosine residue. *Protein Sci* **2018**; 27: 531–545.
(45) Duong VT, Chen Z, Thapa MT, Luo R. Computational Studies of Intrinsically Disordered Proteins. *J Phys Chem B* **2018**; 122: 10455–10469.
(46) Meng F, Bellaiche MMJ, Kim JY *et al.* Highly Disordered Amyloid-β Monomer Probed by Single-Molecule FRET and MD Simulation. *Biophys J* **2018**; 114: 870–884.
(47) Fealey ME, Binder BP, Uversky VN, Hinderliter A, Thomas DD. Structural Impact of Phosphorylation and Dielectric Constant Variation on Synaptotagmin's IDR. *Biophys J* **2018**; 114: 550–561.
(48) Grazioli G, Martin RW, Butts CT. Comparative exploratory analysis of intrinsically disordered protein dynamics using machine learning and network analytic methods. *Front Mol Biosci* **2019**; 6: 42.
(49) Shrestha UR, Juneja P, Zhang Q *et al.* Generation of the configurational ensemble of an intrinsically disordered protein from unbiased molecular dynamics simulation. *Proc Natl Acad Sci U S A* **2019**; 116: 20446–20452.
(50) Robustelli P, Piana S, Shaw DE. The mechanism of coupled folding-upon-binding of an intrinsically disordered protein. *bioRxiv* **2020**. doi:10.1101/2020.03.23.004283.
(51) Wang K, Ning S, Guo Y, Duan M, Yang M. The regulation mechanism of phosphorylation and mutations in intrinsically disordered protein 4E-BP2. *Phys Chem Chem Phys* **2020**; 22: 2938–2948.
(52) Staby L, O'Shea C, Willemoës M *et al.* Eukaryotic transcription factors: Paradigms of protein intrinsic disorder. Biochem. J. **2017**; 474: 2509–2532.
(53) Liu J, Perumal NB, Oldfield CJ *et al.* Intrinsic disorder in transcription factors.

Biochemistry **2006**; 45: 6873–6888.
(54) Nieto MA. The snail superfamily of zinc-finger transcription factors. Nat. Rev. Mol. Cell Biol. **2002**; 3: 155–166.
(55) Qi D, Bergman M, Aihara H, Nibu Y, Mannervik M. Drosophila Ebi mediates Snail-dependent transcriptional repression through HDAC3-induced histone deacetylation. *EMBO J* **2008**; 27: 898–909.
(56) Hemavathy K, Guru SC, Harris J, Chen JD, Ip YT. Human Slug Is a Repressor That Localizes to Sites of Active Transcription. *Mol Cell Biol* **2000**; 20: 5087–5095.
(57) Kanno TYN, Fogo MS, Goes CP, Vieceli FM, Yan CYI. Functional analysis of Scratch2 domains: Implications in the evolution of Snail transcriptional repressors. bioRxiv. **2018**. doi:10.1101/410761.
(58) Huang J, Mackerell AD. Induction of peptide bond dipoles drives cooperative helix formation in the (AAQAA)3 peptide. *Biophys J* **2014**; 107: 991–997.
(59) Sefton M, Sánchez S, Nieto MA. Conserved and divergent roles for members of the Snail family of transcription factors in the chick and mouse embryo. *Development* **1998**; 125: 3111–3121.
(60) Yamaguchi H, Hsu JL, Hung MC. Regulation of ubiquitination-mediated protein degradation by survival kinases in cancer. Front. Oncol. **2012**; 2: 1–9.
(61) Yang DJ, Chung JY, Lee SJ *et al.* Slug, mammalian homologue gene of Drosophila escargot, promotes neuronal-differentiation through suppression of HEB/daughtherless. *Cell Cycle* **2010**; 9: 2861–2874.
(62) Hemavathy K, Ashraf SI, Ip YT. Snail/Slug family of repressors: Slowly going into the fast lane of development and cancer. Gene. **2000**; 257: 1–12.
(63) Voog J, Sandall SL, Hime GR *et al.* Escargot Restricts Niche Cell to Stem Cell Conversion in the Drosophila Testis. *Cell Rep* **2014**; 7: 722–734.
(64) Antonello ZA, Reiff T, Ballesta-Illan E, Dominguez M. Robust intestinal homeostasis relies on cellular plasticity in enteroblasts mediated by miR-8–Escargot switch. *EMBO J* **2015**; 34: 2025–2041.
(65) Hartl M, Loschek LF, Stephan D *et al.* A new prospero and microRNA-279 pathway restricts CO 2 receptor neuron formation. *J Neurosci* **2011**; 31: 15660–15673.
(66) Korzelius J, Naumann SK, Loza-Coll MA *et al.* Escargot maintains stemness and suppresses differentiation in Drosophila intestinal stem cells. *EMBO J* **2014**; 33: 2967–2982.
(67) Yagi Y, Hayashi S. Role of the Drosophila EGF receptor in determination of the dorsoventral domains of escargot expression during primary neurogenesis. *Genes to Cells* **1997**; 2: 41–53.
(68) Sanchez-Díaz I, Rosales-Bravo F, Reyes-Taboada JL *et al.* The Esg Gene Is Involved in Nicotine Sensitivity in Drosophila melanogaster. *PLoS One* **2015**; 10: e0133956.
(69) Hayashi S, Susumu H, Metcalfe T, Shirras AD. Control of imaginal cell development by the escargot gene of Drosophila. *Development* **1993**; 118: 105–115.
(70) Fuse N, Hirose S, Hayashi S. Diploidy of Drosophila imaginal cells is maintained by a transcriptional repressor encoded by escargot. *Genes Dev* **1994**; 8: 2270–2281.
(71) Babu MM, van der Lee R, de Groot NS, Gsponer J. Intrinsically disordered proteins: Regulation and disease. Curr. Opin. Struct. Biol. **2011**; 21: 432–440.
(72) Lai SL, Miller MR, Robinson KJ, Doe CQ. The snail family member worniu is continuously required in neuroblasts to prevent Elav-Induced premature differentiation. *Dev Cell* **2012**; 23: 849–857.
(73) Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search

tool. *J Mol Biol* **1990**; 215: 403–410.
(74) Kozlowski LP, Bujnicki JM. MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. *BMC Bioinformatics* **2012**; 13: 111.
(75) Mizianty MJ, Peng Z, Kurgan L. MFDp2: Accurate predictor of disorder in proteins by fusion of disorder probabilities, content and profiles. *Intrinsically Disord Proteins* **2013**; 1: e24428.
(76) Wang S, Ma J, Xu J. AUCpreD: Proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields. In: *Bioinformatics*. Oxford University Press, 2016, pp i672–i679.
(77) Hanson J, Yang Y, Paliwal K, Zhou Y. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics* **2017**; 33: 685–692.
(78) Nielsen JT, Mulder FAA. Quality and bias of protein disorder predictors. *Sci Rep* **2019**; 9: 5137.
(79) Caswell TA, Droettboom M, Hunter J *et al.* matplotlib/matplotlib v2.2.3. **2018**. doi:10.5281/ZENODO.1343133.
(80) Hunter JD. Matplotlib: A 2D graphics environment. *Comput Sci Eng* **2007**; 9: 99–104.
(81) Bateman A. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res* **2019**; 47: D506–D515.
(82) Agarwala R, Barrett T, Beck J *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **2016**; 44: D7–D19.
(83) Sievers F, Wilm A, Dineen D *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **2011**; 7. doi:10.1038/msb.2011.75.
(84) Holehouse AS, Das RK, Ahad JN, Richardson MOG, Pappu R V. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophys J* **2017**; 112: 16–21.
(85) Millard PS, Bugge K, Marabini R *et al.* IDDomainSpotter: Compositional bias reveals domains in long disordered protein regions—Insights from transcription factors. *Protein Sci* **2019**; : 1–15.
(86) Yang Y, Faraggi E, Zhao H, Zhou Y. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics* **2011**; 27: 2076–2082.
(87) Glykos NM. Software news and updates carma: A molecular dynamics analysis program. *J Comput Chem* **2006**; 27: 1765–1768.
(88) Jo S, Kim T, Iyer VG, Im W. CHARMM-GUI: A web-based graphical user interface for CHARMM. *J Comput Chem* **2008**; 29: 1859–1865.
(89) Söding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* **2005**; 33: W244–W248.
(90) Yang J, Zhang Y. I-TASSER server: New development for protein structure and function predictions. *Nucleic Acids Res* **2015**; 43: W174–W181.
(91) Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **2015**; 10: 845–858.
(92) Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins Struct Funct Bioinforma* **2012**; 80: 1715–1735.
(93) Humphrey W, Dalke A, Schulten K. VMD: Visual Molecular Dynamics. **1996**; 14:

33–38.
(94) Lee J, Cheng X, Swails JM *et al.* CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *J Chem Theory Comput* **2016**; 12: 405–413.
(95) Phillips JC, Braun R, Wang W *et al.* Scalable molecular dynamics with NAMD. J. Comput. Chem. **2005**; 26: 1781–1802.
(96) Best RB, Zhu X, Shim J *et al.* Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone φ, ψ and side-chain χ1 and χ2 Dihedral Angles. *J Chem Theory Comput* **2012**; 8: 3257–3273.
(97) Tanner DE, Chan KY, Phillips JC, Schulten K. Parallel generalized born implicit solvent calculations with NAMD. *J Chem Theory Comput* **2011**; 7: 3635–3642.
(98) Pettersen EF, Goddard TD, Huang CC *et al.* UCSF Chimera - A visualization system for exploratory research and analysis. *J Comput Chem* **2004**; 25: 1605–1612.
(99) Huang J, Rauscher S, Nawrocki G *et al.* CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nat Methods* **2016**; 14: 71–73.
(100) Roe DR, Cheatham TE. PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J Chem Theory Comput* **2013**; 9: 3084–3095.
(101) Abraham MJ, Murtola T, Schulz R *et al.* Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**; 1–2: 19–25.
(102) Hopkins CW, Le Grand S, Walker RC, Roitberg AE. Long-time-step molecular dynamics through hydrogen mass repartitioning. *J Chem Theory Comput* **2015**; 11: 1864–1874.
(103) Balusek C, Hwang H, Lau CH *et al.* Accelerating Membrane Simulations with Hydrogen Mass Repartitioning. *J Chem Theory Comput* **2019**; 15: 4673–4686.
(104) McGibbon RT, Beauchamp KA, Harrigan MP *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys J* **2015**; 109: 1528–1532.
(105) Dohrmann M, Wörheide G. Novel scenarios of early animal evolution-is it time to rewrite textbooks? In: *Integrative and Comparative Biology*. 2013, pp 503–511.
(106) Wägele JW, Bartolomaeus T. *Deep metazoan phylogeny: The backbone of the tree of life: New insights from analyses of molecules, morphology, and theory of data analysis*. Walter de Gruyter GmbH, 2014 doi:10.1515/9783110277524.
(107) Wessel GM. Echinodermata. In: *Encyclopedia of Reproduction*. Elsevier, 2018, pp 533–545.
(108) Tassia MG, Cannon JT, Konikoff CE *et al.* The global diversity of Hemichordata. PLoS One. **2016**; 11. doi:10.1371/journal.pone.0162564.
(109) Rychel AL, Smith SE, Shimamoto HT, Swalla BJ. Evolution and development of the chordates: Collagen and pharyngeal cartilage. Mol. Biol. Evol. **2006**; 23: 541–549.
(110) Kyte J, Doolittle RF. A Simple Method for Displaying the Hydropathic Character of a Protein. 1982.
(111) Sethi A, Tian J, Vu DM, Gnanakaran S. Identification of minimally interacting modules in an intrinsically disordered protein. *Biophys J* **2012**; 103: 748–757.
(112) Lee KH, Chen J. Optimization of the GBMV2 implicit solvent force field for accurate simulation of protein conformational equilibria. *J Comput Chem* **2017**; 38: 1332–1341.
(113) Chen J, Im W, Brooks CL. Balancing solvation and intramolecular interactions:

Toward a consistent generalized born force field. *J Am Chem Soc* **2006**; 128: 3728–3736.
(114) Best RB, Hummer G. Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *J Phys Chem B* **2009**; 113: 9004–9015.
(115) Best RB, Mittal J, Feig M, MacKerell AD. Inclusion of many-body effects in the additive CHARMM protein CMAP potential results in enhanced cooperativity of α-helix and β-hairpin formation. *Biophys J* **2012**; 103: 1045–1051.
(116) Robustelli P, Piana S, Shaw DE. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc Natl Acad Sci U S A* **2018**; 115: E4758–E4766.
(117) Liu H, Song D, Zhang Y *et al.* Extensive tests and evaluation of the CHARMM36IDPSFF force field for intrinsically disordered proteins and folded proteins. *Phys Chem Chem Phys* **2019**; 21: 21918–21931.
(118) Tomasso ME, Tarver MJ, Devarajan D, Whitten ST. Hydrodynamic Radii of Intrinsically Disordered Proteins Determined from Experimental Polyproline II Propensities. *PLoS Comput Biol* **2016**; 12: e1004686.
(119) Song D, Luo R, Chen HF. The IDP-Specific Force Field ff14IDPSFF Improves the Conformer Sampling of Intrinsically Disordered Proteins. *J Chem Inf Model* **2017**; 57: 1166–1178.
(120) Das RK, Pappu R V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc Natl Acad Sci U S A* **2013**; 110: 13392–13397.
(121) Basu S, Biswas P. Salt-bridge dynamics in intrinsically disordered proteins: A trade-off between electrostatic interactions and structural flexibility. *Biochim Biophys Acta - Proteins Proteomics* **2018**; 1866: 624–641.
(122) Donald JE, Kulp DW, DeGrado WF. Salt bridges: Geometrically specific, designable interactions. *Proteins Struct Funct Bioinforma* **2011**; 79: 898–915.
(123) Zhang X, Li L, Li N *et al.* Salt bridge interactions within the β2 integrin α7 helix mediate force-induced binding and shear resistance ability. *FEBS J* **2018**; 285: 261–274.
(124) Basu S, Mukharjee D. Salt-bridge networks within globular and disordered proteins: characterizing trends for designable interactions. *J Mol Model* **2017**; 23: 206.
(125) Huerta-Viga A, Amirjalayer S, Domingos SR *et al.* The structure of salt bridges between Arg+ and Glu- in peptides investigated with 2D-IR spectroscopy: Evidence for two distinct hydrogen-bond geometries. *J Chem Phys* **2015**; 142: 212444.
(126) Vijayakumar M, Qian H, Zhou HX. Hydrogen bonds between short polar side chains and peptide backbone: Prevalence in proteins and effects on helix-forming propensities. *Proteins Struct Funct Genet* **1999**; 34: 497–507.
(127) Wang J, Feng JA. Exploring the sequence patterns in the α-helices of proteins. *Protein Eng* **2003**; 16: 799–807.
(128) Mateos B, Conrad-Billroth C, Schiavina M *et al.* The Ambivalent Role of Proline Residues in an Intrinsically Disordered Protein: From Disorder Promoters to Compaction Facilitators. *J Mol Biol* **2019**; 432: 3093-3111.
(129) Vymětal J, Bednárová L, Vondrášek J. Effect of TFE on the Helical Content of AK17 and HAL-1 Peptides: Theoretical Insights into the Mechanism of Helix Stabilization. *J Phys Chem B* **2016**; 120: 1048–1059.
(130) Mouillon JM, Gustafsson P, Harryson P. Structural investigation of disordered stress proteins. Comparison of full-length dehydrins with isolated peptides of their conserved segments. *Plant Physiol* **2006**; 141: 638–650.

ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

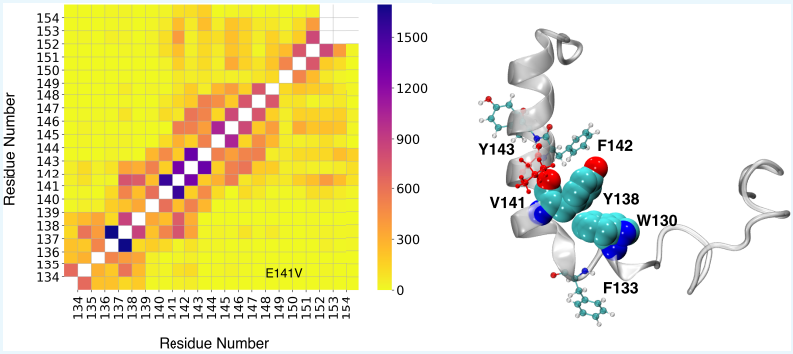
Identification of an α-MoRF in the Intrinsically Disordered Region of the Escargot Transcription Factor

Teresa Hernández-Segura and Nina Pastor*

Cite This: *ACS Omega* 2020, 5, 18331−18341

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



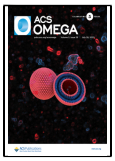
ABSTRACT: Molecular recognition features (MoRFs) are common in intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs). MoRFs are in constant order−disorder structural transitions and adopt well-defined structures once they are bound to their targets. Here, we study Escargot (Esg), a transcription factor in *Drosophila melanogaster* that regulates multiple cellular functions, and consists of a disordered N-terminal domain and a group of zinc fingers at its C-terminal domain. We analyzed the N-terminal domain of Esg with disorder predictors and identified a region of 45 amino acids with high probability to form ordered structures, which we named S2. Through 54 μs of molecular dynamics (MD) simulations using CHARMM36 and implicit solvent (generalized Born/surface area (GBSA)), we characterized the conformational landscape of S2 and found an α-MoRF of ∼16 amino acids stabilized by key contacts within the helix. To test the importance of these contacts in the stability of the α-MoRF, we evaluated the effect of point mutations that would impair these interactions, running 24 μs of MD for each mutation. The mutations had mild effects on the MoRF, and in some cases, led to gain of residual structure through long-range contacts of the α-MoRF and the rest of the S2 region. As this could be an effect of the force field and solvent model we used, we benchmarked our simulation protocol by carrying out 32 μs of MD for the $(AAQAA)_3$ peptide. The results of the benchmark indicate that the global amount of helix in shorter peptides like $(AAQAA)_3$ is reasonably predicted. Careful analysis of the runs of S2 and its mutants suggests that the mutation to hydrophobic residues may have nucleated long-range hydrophobic and aromatic interactions that stabilize the MoRF. Finally, we have identified a set of residues that stabilize an α-MoRF in a region still without functional annotations in Esg.

1. INTRODUCTION

Intrinsically disordered proteins (IDPs) or intrinsically disordered regions (IDRs) have no stable, well-defined three-dimensional structures under physiological conditions. They play an important role in many biological functions, such as organ development, and their dysfunction is associated with multiple diseases.[1−3] One of the main goals of studying IDPs is to understand their capacity to adopt multiple conformations and to explore the binding mechanisms leading to interactions and their biological functions.

For those IDPs that fold upon binding, this has been analyzed in terms of two extreme models: "conformational selection" and "induced fit".[4] In the conformational selection mechanism, the protein explores conformations that are both ordered and disordered, and specific folded conformations are selected by the ligand; in the induced fit mechanism, the conformational change occurs after the disordered protein is bound to the ligand[4,5] and this process is known as "template folding", where the folding transition state is dictated by the interactions with the ligand.[6] Real cases present a mixture of these two scenarios.

Received: May 3, 2020
Accepted: July 2, 2020
Published: July 17, 2020



https://dx.doi.org/10.1021/acsomega.0c02051
ACS Omega 2020, 5, 18331−18341

Residual structure often persists in unbound IDPs.[7−9] These preformed but unstable structural elements might serve as initial contact points to facilitate the folding and recognition of the ligand surface.[7−10] These residual structures may exist in short regions known as molecular recognition features (MoRFs), which can be found in long disordered regions of IDPs and IDRs.[11,12] MoRFs are mainly disordered in their unbound states and adopt local structures that are stabilized when they interact with their targets.[11] MoRFs are classified according to their structures in the bound state, where α-MoRFs form α-helices, β-MoRFs form β-strands, ι-MoRFs form irregular structures (coil conformations), and complex MoRFs adopt a conformation resulting from the combination of the other types.[13,14] MoRFs play important roles in signaling and regulatory functions,[12,15] and transcription factors are enriched in IDRs.[16,17]

Exploring and characterizing the conformational ensemble of IDPs is not an easy task, as they are represented by multiple conformations, and different ones can be associated with particular functions. Many experimental techniques, such as dynamic light scattering, small-angle X-ray scattering, paramagnetic relaxation enhancement, and circular dichroism, are useful to obtain information on the ensemble-average of diverse conformations of IDPs. Computational techniques are also used; for example, molecular dynamics (MD) and Monte Carlo (MC) simulations have been useful to complement the information obtained by experimental techniques and to characterize the conformational ensembles at the atomic level.[1,2,18,19]

Simulations allow us to study the molecular features and motions of IDPs. However, structural characterization by simulations is limited by the long time required for adequate conformational sampling and by the accuracy of the parameters of the force fields and the explicit or implicit solvent models used. Protein force fields were developed and parameterized mainly to describe folded proteins[2,20] and their application has been thoroughly benchmarked, but their applicability to IDPs requires considerable attention and continuous improvement;[1,19] different force fields and protocols with explicit and implicit solvent models have been applied to various IDPs (summarized in Table S1 for the past 5 years).

The motions of conformational changes in an IDP happen in time scales spanning several orders of magnitude (between picosecond (ps) to microsecond (ms))[21] and this implies substantial computational costs, depending on the system size.[22] The behavior of protein chains in solution depends fundamentally on the balance of solute−solute versus solute−solvent and solvent−solvent interactions.[23] The effects of the solvent environment can be treated using explicit or implicit solvent representations.[8,23] Explicit solvents treat both protein and water molecules explicitly, significantly increasing the system size (∼10 fold) and leading to large computer requirements[24] in simulations of protein folding[8,23,25] or characterization of the conformational ensembles of IDPs.[25] This approach has been successful in reproducing and explaining experimental measurements.[3,26−29] Alternatively, implicit solvent models are popular because they include explicitly only the atomic coordinates of the protein, as the water molecules are represented by an infinite continuum medium with the macroscopic properties of water; this reduces the computational cost and allows for longer simulation times required for simulating conformational transitions, especially in novo simulations of IDPs.[7,23,24,30] Modeling small IDPs with an implicit solvent has also been successful.[23,25,31−33]

However, both explicit and implicit solvent models have limitations. For example, some combinations of force fields and water models such as TIP3P, TIP4P, and TIP4P-EW generate overly compact conformational IDP ensembles.[2,34,35] As some IDPs are enriched in charged and polar residues,[1,36] their structural properties make them sensitive to interactions with water.[2] This effect of overcompaction could be the result of underestimating water−water and water−protein interactions[34] or a mismatch between the protein force field and the water model.[35] For this reason, it is paramount to validate the combination of the force field and water model used.[20] On the other hand, many implicit solvent models also tend to generate conformations that are too structured and compact.[22,37] There are important structural properties of water and short-range effects that are not considered in implicit solvent models, and their disregard can lead to different folding mechanisms.[38,39] Optimization efforts have also led to improvements in the implicit solvent models,[23] such as the ABSINTH implicit solvent force field that was developed and optimized specifically for IDPs.[8,31]

Some of the force fields generate more compact conformations than others, as well as overstabilization of specific conformations.[20] Nowadays, there is no standard, specific protocol to simulate IDPs, but there is a continuous search for developing and improving force fields and sampling strategies that can be applied to both folded proteins and IDPs.[19] As can be seen from the simulations listed in Table S1, both MD and MC simulations are useful, and enhanced sampling techniques such as replica exchange (the temperature and the Hamiltonian versions) have also been used. Given the vastness of the conformational space to the sample, another strategy is to guide the simulations with experimental data or to reweigh the simulated ensemble to approximate the experimental data. These procedures only work for original ensembles that are close to the target, so the quality of the force fields and sampling must be good. Fortunately, significant progress has recently been made in the ability of MD force fields to accurately describe folded proteins and IDPs.[40−43] Central to this progress is having small benchmark systems that are well characterized experimentally, such as the 15-residue $(AAQAA)_3$ peptide that has been used for studying protein folding, characterize the helix−coil transitions, and to parameterize and validate different force fields[21,38,39,41,42] (summarized in Table S2).

Here, we study the N-terminal disordered domain of Escargot (Esg), a transcription factor of the Snail family, to look for potential MoRFs associated with Esg functions that are independent of its DNA-binding activity. We performed MD simulations starting from semifolded structures to characterize the conformational ensemble of a region with high propensity to order, using CHARMM36 and an implicit solvent model (generalized Born model, or generalized Born/surface area (GBSA)). Having identified a probable residual structure that adopts α-MoRF conformations and detailed the interactions that stabilize it, we proposed mutations that would destabilize it and simulated them as well. Contrary to our expectations, the mutations did not lead to dramatic losses of structure. Careful analysis of the inherent bias of our simulation protocol and of the interactions that stabilize the residual structure in the mutants revealed a network of hydrophobic and aromatic interactions stabilizing the α-MoRF.

https://dx.doi.org/10.1021/acsomega.0c02051
ACS Omega 2020, 5, 18331−18341

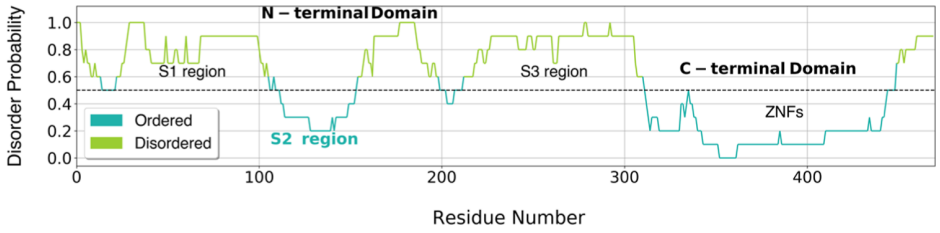


Figure 1. Disorder probability analysis of the Esg protein. The plot shows the average probability of disorder by residue. The light and dark green color lines show the disordered and ordered regions, respectively.

This allows us to propose a set of critical residues for the formation and stability of this α-MoRF; these could be tested by mutagenesis experimentally, to determine the functional role of this region of Esg, which lacks annotated functional motifs. Our results might also be helpful as a benchmark for this combination of a force field and a solvent model, aiding in future refinements of both. This is the first structural analysis of the N-terminal region of Esg and aims to foster future studies into the structure–function relationship of this region to understand the mechanism of regulation of this transcription factor.

2. RESULTS AND DISCUSSION

Esg is a protein of 470 residues, which is expressed in *Drosophila melanogaster*, and is involved in the development of the nervous system.[44,45] Structurally, the Esg C-terminal domain (CTD) is a conserved region that has five classical zinc fingers (ZNFs) and interacts with nucleic acids,[46] while the N-terminal domain is divergent,[45,46] and the only functional annotation in this region consists of two motifs (P-DLS-K) that interact with the C-terminal binding protein (CtBP),[47] a transcriptional repressor. However, the N-terminal domain (NTD) has been associated with functions such as protein degradation, where the ZNFs are not necessary,[44] but the actual functional motifs for this and other activities have not been described.

2.1. The NTD of Esg is an IDR. Using four disorder predictor programs and averaging the output scores of all of the predictors, we generated a profile of disorder (Figure 1) that reveals that the NTD is highly disordered, in contrast to the CTD (residues 310–470) containing the ZNFs. The disorder profile of the NTD was split into three regions: S1 (residues 1–110), S2 (residues 111–155), and S3 (residues 156–309), where the S2 region has a much higher probability to adopt order than the other two.

2.2. Structural Predictions for the S2 Region. Currently, there are no known structures of homologues of this region of Esg. Therefore, we used five different structure predictors to obtain structural models for the S2 region (Figure 2). The predictions consistently include an α-helix in the C-terminus of the region, sometimes together with a β-hairpin in the N-terminus, except for a model, which suggests that the S2 region adopts a β-sheet structure (Figure 2, SP6 model). The results of these predictors have low confidence scores because of the lack of homologues with known structures, so on their own, they should not be used as representative structures of S2. Nevertheless, they are useful as initial coordinates for the MD simulations, and to reveal secondary structure preferences inherent to the amino acid sequence. The conformational diversity of these 27 structures is documented in Table S3, a

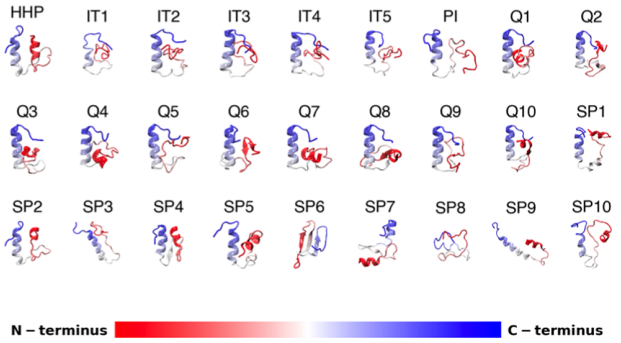


Figure 2. Initial ensemble of the S2 region in the NTD of Esg, obtained by HHpred (HHP), I-Tasser (IT), Phyre2 (PI), QUARK (Q), and SPARK-X (SP) predictors. Each model shows secondary structural elements and is color-coded from red to blue, progressing from the N- to the C-terminus.

compilation of the pairwise root-mean-square deviation (RMSD) between all of them; these range from 3 to 14.1 Å.

The IDPs and IDRs are represented as a dynamic ensemble, which is characterized by different conformations. MD simulations can be used to generate many conformations of the IDP and characterize its conformational ensemble through structural and dynamical information.[3,21] In this work, we have considered running a collection of short (2 μs) simulations using a set of 27 starting models rather than one long simulation, to explore the S2 region conformational ensemble; this strategy has been shown to give more efficient conformational sampling and increases the probability of converging to experimental data.[48]

2.3. Conformational Sampling of the S2 Region. A serious concern with MD simulations is their degree of convergence, and whether the simulation time has been enough to explore the properties of interest. As a first rough measure of the sampling achieved in the 2 μs runs, we calculated the distribution of the C_α-atoms' root-mean-square deviation (RMSD) to the initial structure of each run to ascertain how much they had wandered from the starting point, and the radius of gyration (R_g), as a measure of compaction.[49,50] An energy landscape built with RMSD and R_g for the 54 μs ensemble (Figure 3A) shows that the conformational distribution is located at two basins, one with low RMSDs and small R_gs, and a larger one with greater variation both in RMSD and R_g. The individual histograms of RMSD and R_g are shown in Figure 3B,C, respectively. The RMSD showed five evident peaks centered near 1.5, 6, 8.5, 9.5, and 13.5 Å, indicating that some runs remained very close to their starting point, while others roamed more freely (Figure 3B); low RMSD happens with low R_g, as expected because more intramolecular contacts hinder conformational explora-

https://dx.doi.org/10.1021/acsomega.0c02051
ACS Omega 2020, 5, 18331–18341

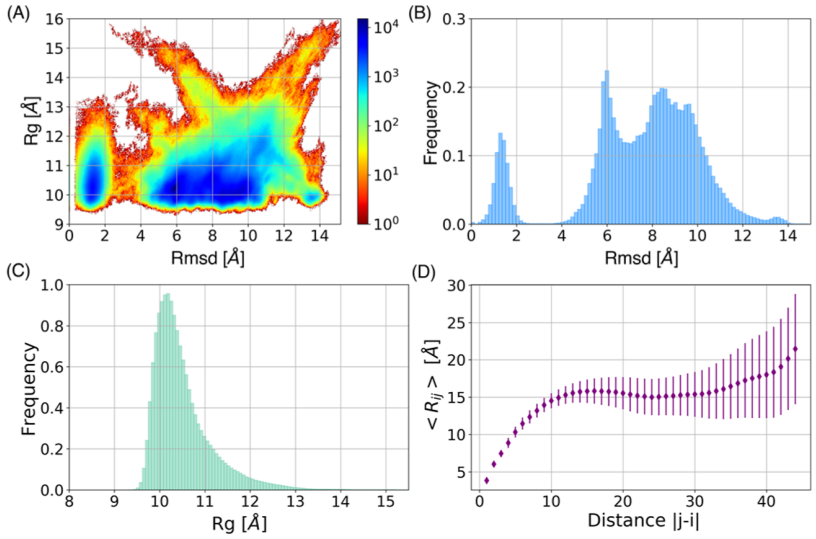


Figure 3. Structural diversity and degree of compaction of the S2 region of Esg. (A) Energy landscape built with C_α-atom RMSD (Å) and R_g (Å). (B) Histogram of the C_α-atom RMSD (Å) distribution in the ensemble during 54 μs of simulation. (C) Histogram of the R_g (Å) distribution in the ensemble during 54 μs of simulation. (D) Average inter-residue distances (Å) as a function of sequence distance during 54 μs of simulation.

tion. Considering the properties of the amino acid sequences of IDPs, it is possible to estimate their hydrodynamic radii (R_h).[51] For the S2 region with 45 residues, the theoretically estimated R_h value is around 15 Å, and considering that the diameter of a molecule of water is ~3 Å, the corresponding R_g should be ~12 Å. The distribution of R_g showed one single peak located between 10 and 11 Å, which corresponds to semicompact conformations (Figure 3C). The results showed that the S2 region ensemble could adopt flexible and heterogeneous conformations, characteristic of an IDP.[3] Some force fields and implicit solvent models generate overestimation of compactness for the IDP ensembles,[52,53] and Figure 3 indicates that our simulations also show a modest increase in compaction compared to the predicted R_g. However, chain compaction in IDPs depends on their sequences, for example, the fraction of charged residues and proline content.[54,55] Both R_g and R_h have been related to net charge per residue (NCPR),[54] where IDPs with NCPR > 0.25 adopt expanded-coil conformations, while NCPR < 0.25 indicates compact globular conformations. The S2 region of Esg has three charged residues, a net charge of +1, 21 hydrophobic, 14 polar, seven aromatic, and six proline residues, and 22 of its 45 amino acids are promoters of disorder. It has an NCPR = 0.022 obtained by the CIDER server,[56] so it is expected to adopt compact conformations. Figure 3D shows that the average inter-residue distances are larger than those expected for a Lennard-Jones collapsed structure,[53,54] but much smaller than those of a Flory chain of the same length. At this point, it should be stressed that there is no experimental data for S2 that we could use to guide our simulations or as a litmus test of the quality of the data set.

One of the advantages of running multiple independent simulations starting from different structures is that this accelerates convergence in principle. To determine whether 54 μs of simulation is an adequate simulation time, we looked for structurally similar conformations sampled by two different runs. A matrix was generated by the pairwise RMSD between pairs of trajectories, comparing each generated structure in one run to each of the other. Figure 4 shows the heat maps with the pairwise RMSD distance computed between C_α-atoms of the

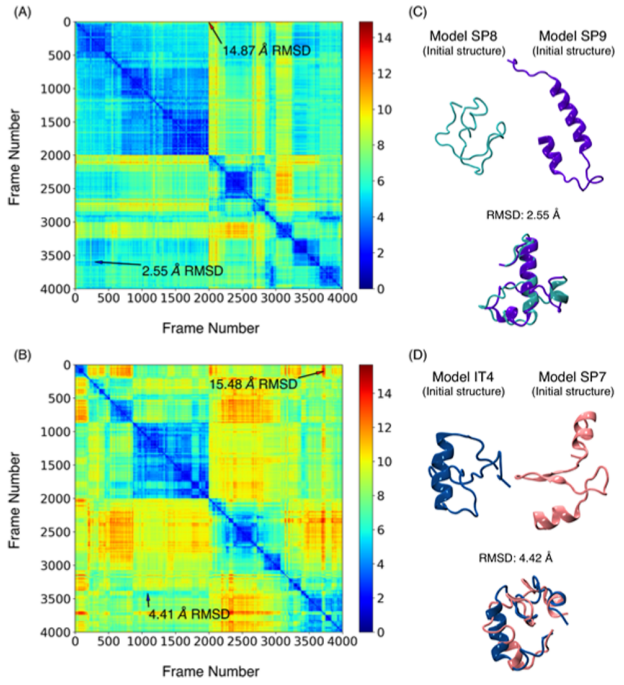


Figure 4. Heat maps representing pairwise RMSD (Å) calculated for the C_α-atoms of the (A) SP8 (frames 1−2000) and SP9 (frames 2001−4000) runs and (B) IT4 (frames 1−2000) and SP7 (frames 2001−4000) runs. Each plot shows the location of the minimum and maximum pairwise RMSD. (C) Initial structures of the SP8 and SP9 models, and overlapping of the minimum of pairwise RMSD. (D) Initial structures of the IT4 and SP7 models, and overlapping of the minimum of pairwise RMSD.

structures generated in the SP8 and SP9 model runs (Figure 4A), and IT4 and SP7 model runs (Figure 4B). Pairs of structures with the smallest pairwise RMSD are in blue, while the largest RMSD is in dark red. The diagonal dark blue line represents the pairwise RMSD comparison of a structure with itself, and the two diagonal squares (frame numbers 1−2000 and 2001−4000) correspond to the comparison within each of

https://dx.doi.org/10.1021/acsomega.0c02051
ACS Omega 2020, 5, 18331−18341

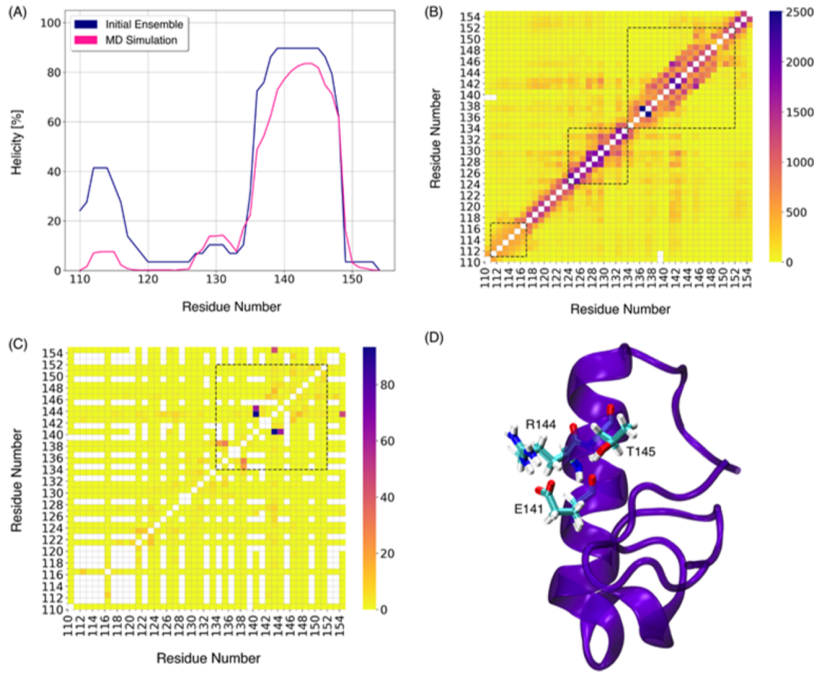


Figure 5. Secondary structure and tertiary contacts of the S2 region. (A) Percentage of the time found as a helix for each residue. (B) Heat map representing the contacts between residue pairs; the dashed line square marks the initial position of helices. (C) Heat map representing the hydrogen bonds involving side chain atoms between residue pairs; the dashed line square marks the C-terminal helix. (D) α-Helix conformation depicting the interactions between E141 with R144 and T145. The main chain is shown as a purple ribbon, and the amino acids are shown as sticks in CPK colors.

the runs; the interesting areas of these plots are the off-diagonal squares, where one run is compared to the other. The pairwise RMSD for each model shows the transition between conformational states within one run.

SP8 and SP9 started with a pairwise RMSD of 14.1 Å (Table S3), and during their runs, they generated structures that drifted further apart (maximum pairwise RMSD of 14.87 Å) and also became similar to each other, reaching a pairwise RMSD minimum of 2.55 Å; it can also be seen that low RMSDs occurred multiple times during these two runs (Figure 4A). Similar results were obtained with the IT4 and SP7 trajectories (Figure 4B), where the minimum pairwise RMSD was 4.42 Å and the maximum was 15.48 Å, having started at 10.9 Å. The overlap of the two structures with the minimum pairwise RMSD is shown in Figure 4C,D, illustrating the degree of conformational convergence. This result suggests that simulating each model for 2 μs gives a reasonable run time, so we assume that the structures generated by the 54 μs of MD simulation are a good starting approximation to the diverse conformational space of the S2 region of Esg. A compilation of minimum pairwise RMSDs between all of the models can be found in Table S4; none exceeds 8 Å.

2.4. The S2 Region Harbors an α-MoRF. Inspection of the structures in Figure 2 reveals a prevalence of α-helices. The secondary structure propensity for each amino acid, contained in the structure predictors, yielded two helices: one near the N-terminus in the region from residues 111 to 120 and a second one in the C-terminus, spanning residues 125−150 (Figure 5A). During the simulation, the secondary structure content is lower than that in the initial ensemble; the α-helix conformation of the N-terminus was decreased by ∼30%, while the α-helix conformation of the C-terminus was decreased by ∼10% (Figure 5A). Previous MD simulations have reported that protein unfolding can occur on a picosecond time scale[49] and it is thus interesting that during 54 μs of simulations, the α-helix conformation remains the most persistent secondary structure in the C-terminus of S2.

Persistent structures require stabilizing interactions. To find those, the total numbers of contacts and hydrogen bonds along the simulations were computed and are presented in the heat maps of Figure 5B,C, respectively. The frequency of interaction between residue pairs is indicated by the color bar, where the dark blue color indicates those contacts with the highest frequency and the yellow color indicates those with the least frequency; white squares away from the diagonal indicate that no interactions were found between that pair of residues. Due to the flexibility of the IDPs, many transient long-range contacts are expected, and these can be seen in the yellow regions. A large probability of contacts between nearby residues in the primary sequence is a hallmark of helical structures,[57] and these are identified with the black squares of dashed lines in Figure 5B,C.

In Figure 5B,C, the diagonal line represents the interaction of a residue with itself; short-range interactions between residues lie close to the diagonal, and these indicate helices. Off-diagonal interactions are considered long range, and represent tertiary interactions and/or the formation of hairpins. The initial N-terminal α-helices (between residues 111−117 and 124−134) are substantially weakened during the simulations. The region of highest helicity is present in the C-terminal of the S2 region, between the residue regions from 134 to 152, where the most frequent inter-residue contacts and hydrogen bonds occur. Hydrogen bonds contribute to the stability of the protein secondary structure.[49] The α-helix conformation could be stabilized by interactions between E141 and R144 and T145 residues, which are in the middle of the α-

https://dx.doi.org/10.1021/acsomega.0c02051
ACS Omega 2020, 5, 18331−18341

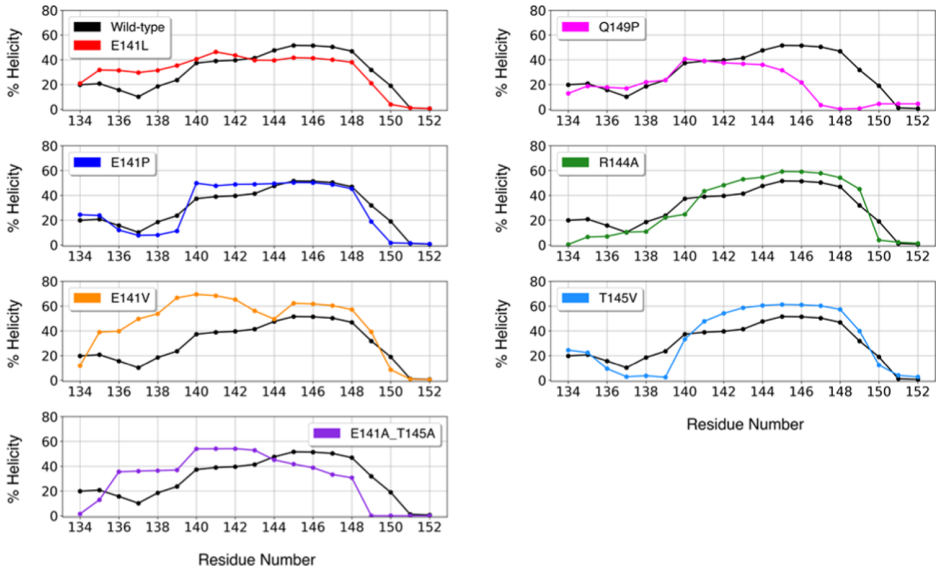


Figure 6. Percentage of the helicity per residue of the α-MoRF (from residues 134 to 152) of each mutant, with respect to the wild-type (black line). The value for each residue is marked by a dot.

helix (Figure 5D), and are the most frequent side chain hydrogen bond interactions in the ensemble. Together, these results suggest that the S2 region is an IDR and that it could harbor an α-MoRF.

2.5. Mutations to Disrupt the α-MoRF. Oppositely charged residues lead to the formation of salt bridges,[36,58] stabilizing local structures that can be relevant in molecular recognition events.[57−61] Salt bridges are known to impart local rigidity and the disruption of these interactions increases flexibility,[36,59] which could be associated with "order to disorder" structural transitions in IDPs.[36] We have found that the most persistent hydrogen bonds in the ensemble are those between E141, R144, and T145. Interactions between E and R contribute favorably to the stability of helices,[60,61] and interactions between E and R (i + 3, i + 4) are more favorable than the reversed salt bridge.[61] Glutamate can establish side chain−backbone and side chain−side chain interactions,[62] while threonine has a short polar side chain and has a strong tendency to form hydrogen bonds with neighboring backbone amides, as seen in our simulations. Side chain−backbone hydrogen bonds are determinant to helix formation and the elimination of these interactions destabilizes the helix.[62] We therefore mutated these residues to eliminate the interactions and disrupt the α-MoRF. E141, R144, and T145 were substituted by residues that cannot form hydrogen bonds and/or have lower helical propensity, such as valine and proline.[63] E141 was substituted by leucine (E141L), valine (E141V), and proline (E141P). A double mutant (E141A_T145A) was generated, to eliminate both side chain−side chain and side chain−main chain interactions between E141 and T145. T145 was also substituted by valine (T145V) and R144 by alanine (R144A). Finally, Q149P is an Esg variant in the S2 region that does not show phenotype,[64] and it was also simulated as a control. We selected four models of the initial ensemble (HHP, PI, SP6, and SP9) and generated the mutants by side chain substitution. Each model was simulated during 6 μs, for a total 24 μs ensemble of each wild-type and mutant S2. Figure 6 shows the secondary structure of the α-MoRF region (from residues 134 to 152) of each mutant compared to that in the wild-type. The mutants showed either modest changes or local increases of helicity with respect to the wild-type, toward the N-terminus or the C-terminus of the helix. The Q149P and the double mutant disrupted the C-terminus of the α-MoRF; the loss of structure for Q149P might indicate that the α-MoRF need not extend that far to be functional. The breaking of intramolecular hydrogen bonds has previously been identified as a key step in protein unfolding,[49] but the hydrogen bonds involving side chain atoms in the mutants did not show significant differences compared to the wild-type, and their frequency was very low. On the other hand, the numbers of carbon−carbon contacts increased, especially between the α-MoRF region and the rest of S2; these contacts were not present significantly in the wild-type variant (Table S5).

The mutants have interactions in common, involving aromatic residues on both the helix (Y138, F142, and Y143, surrounding the mutated residues) and outside of it (Y125, Y128, W130, and F133), which stabilize the helical structure (Figure S1 and Table S5) by the possible formation of a small hydrophobic core. The S2 region is enriched in aromatic residues and prolines that could be involved also in the stabilization of the α-MoRF. Recently, it was reported that interactions between prolines and aromatic residues in positions $i \pm 1$ and $i - 2$ can form incipient structural nucleation sites in IDPs.[65] The frequency of these interactions in the simulations is compiled in Table S6 and representative structures depicting these are included in Figure S2. These interactions are common in the simulations and are notably enriched for the E141A_T145A double mutant in all proline−aromatic pairs, and approximately one helix turn away from the mutation site in the E141P variant. All of the proline residues in the simulations are in the trans configuration. To explore biases in their conformational sampling, we calculated the Ramachandran plot for the wild-type 54 μs ensemble (Figure S3), which shows populations of both the polyproline-II and α helix basins, with a small preference for the latter.

2.6. Validation of the Force Field and Solvent Model Used in the Simulations. A well-known issue in the

https://dx.doi.org/10.1021/acsomega.0c02051
ACS Omega 2020, 5, 18331−18341

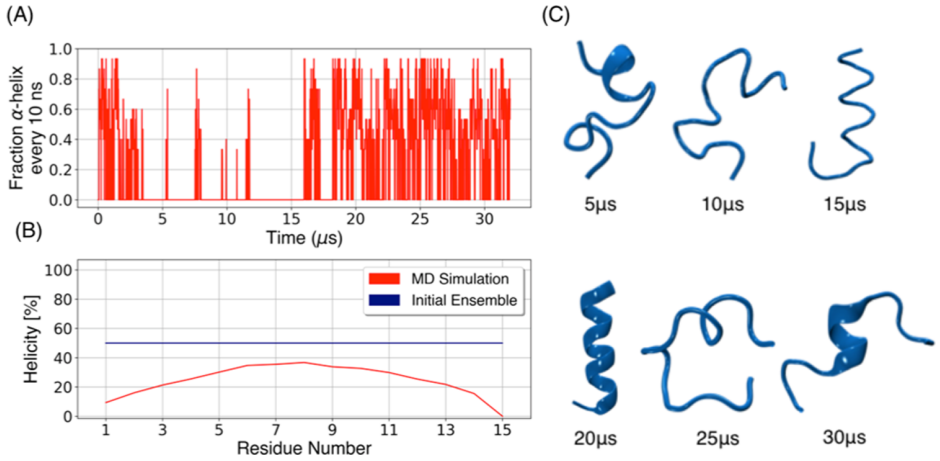


Figure 7. Ensemble of $(AAQAA)_3$ during 32 μs of MD simulation. (A) Fraction helix of $(AAQAA)_3$ computed over 10 ns blocks; the first 16 μs correspond to the simulation starting from the extended conformation and the last 16 μs correspond to the simulation starting from a perfect helix. (B) Percentage of the helicity per residue, averaged over the 32 μs (red line) compared to the initial ensemble (blue line). (C) $(AAQAA)_3$ conformations representing coil−helix transitions at different times during the simulations.

simulation of IDPs is the artificial increase in both secondary and tertiary structures. A modified version of CHARMM36 (CHARMM36m) corrects the excessive population of the left-handed α-helix (α_L-helix) conformation generated in simulations with CHARMM36.[42] This new release was not available when this project started, so all of the runs were carried out with the same force field for consistency. To determine the level of inaccuracy of our description of S2, we calculated the α_L-helix population of the 54 μs of the S2 region ensemble, obtaining a 1.36% population, which is significantly lower than that found in other simulations generated with CHARMM36 of IDPs (between 5.7 and 20%), and considered it within the margin of error of the experimental value.[42,66]

One of the standard peptides used for the validation of force fields is $(AAQAA)_3$, a helical peptide that has been well studied and characterized experimentally to understand the helix−coil transitions.[39,42,66] $(AAQAA)_3$ has ∼19−21% helical content.[42,66] We carried out two 16 μs of simulations of $(AAQAA)_3$ using exactly the same protocol as that for the S2 runs. This particular combination of CHARMM36 with the GBSA solvent (as implemented in NAMD) has not been benchmarked with $(AAQAA)_3$, despite having been used to simulate other IDPs, as shown in Tables S1 and S2. One of the simulations started from a perfect helix (100% helicity) and the other from a completely extended conformation (0% helicity), therefore accounting for the 50% helicity per residue for the starting structures. Figure 7 shows that $(AAQAA)_3$ exchanges frequently between coil and helical conformations (Figure 7A,C), and its average helical content during the 32 μs ensemble was ∼24%, indicating that CHARMM36 with GBSA provides a reasonable balance between helix and coil conformations for $(AAQAA)_3$. Close inspection of Figure 7A shows that the two 16 μs runs had a different behavior: one with an excess of helix and the other with long periods of absence of helices. This indicates that even for simple systems like this small peptide, having seen multiple helix−coil transitions does not guarantee sufficient sampling, so at least two runs starting from different structures should be made of the same system.

Regarding the stabilization of the S2 mutants described above, there remains an unexplored issue for which we have no adequate control or experimental information for contrast. Given that the helices appear to be stabilized by aromatic interactions, this could be due to a real nucleating event caused by increasing locally the hydrophobicity of the helix, as all of the mutations changed a charged or polar residue for a hydrophobic one. Alternatively, this could reflect an imbalance in the polar and nonpolar surface descriptions in the GBSA model. To explore the possibility of indiscriminate hydrophobic collapse, we calculated the contact map for the 32 μs simulation of $(AAQAA)_3$, shown in Figure S4; the structures depicting the most common interactions are included in Figure S5, and, as expected because of their greater number of carbon atoms, involve the glutamine residues. It is clear from this contact map that no long-range contacts are promoted in this peptide, which is 80% hydrophobic. Therefore, we are currently inclined to ascribe the stabilization of the S2 mutants to bona fide interactions, as the simulations of the wild-type and mutant S2 started from the same four conformations and were run with the same protocol, for three times as long as the time we determined was enough to see similar conformations sampled by different runs (see Figure 4).

3. CONCLUSIONS

In this work, we investigated the presence of MoRFs in the NTD of the transcription factor Esg by MD simulations using the CHARMM36 force field and GBSA as an implicit solvent model. In summary, we have detected a region with a high probability to be ordered that we have called the S2 region (from residues 111 to 155). The conformational ensemble of this region contains residual structure as an α-MoRF that persists during 54 μs of MD simulation. In addition, point mutations were built to disrupt this putative α-MoRF and probe its stability during 24 μs of simulation. However, some of the mutants showed an increase of helicity at the α-MoRF, as a consequence of an increase in long-distance contacts with residues outside the MoRF. To validate our simulation protocol, we have evaluated the propensity of the α_L-helix in the 54 μs of MD simulation, and we found that it is significantly lower than that found in ensembles generated with CHARMM36 for other IDPs and in agreement with experimental data. Also, we carried out 32 μs of simulation

https://dx.doi.org/10.1021/acsomega.0c02051
ACS Omega 2020, 5, 18331−18341

of $(AAQAA)_3$ and its average helical content was ∼24%, a value close to the ∼20% measured experimentally, indicating that CHARMM36 together with GBSA provides a good balance between helix and coil conformations for this system. The lack of long-range contacts in these simulations suggests that there is a reasonable balance in the description of polar and nonpolar interactions in this particular combination of a protein force field and an implicit solvent model. Therefore, the stabilization of the MoRF that we found with the mutants could be due to the increase in hydrophobicity at the surface of the helix, which nucleates the interaction with a group of aromatic residues within and outside the MoRF. Assuming that the interactions are correctly described, as we see no such clustering in the wild-type S2 or in the control peptide, we can now propose a set of mutations that should eliminate the MoRF: E141P as a helix breaker, and W130A and F142A as disruptors of the most common long-range interaction. These can be tested in in vitro assays as peptides, or in the context of the full protein in in vivo studies, contributing to the functional annotation of a poorly characterized domain in an important transcription factor, conserved in all multicellular animals.

4. METHODS

4.1. Disorder Analysis. The sequence of the full Esg protein (470 amino acids) was obtained from Uniprot,[67] (Uniprot ID: P25932, htttp://www.uniprot.org/). Disorder analysis was performed using the disorder predictors: MetaDisorder,[68] MFDp2,[69] AUCPred,[70] and SPOT-disorder,[71] which yield the best results as recently reviewed.[72] Metadisorder is a metaserver that utilizes different disorder predictors and calculates an average of the output of those results, thus improving the prediction accuracy. The output of the three variants of Metadisorder was used: MetaDisorderMD, MetaDisorderMD2 (a variant of MetaDisorderMD with a different scoring function), and MetaDisorderMD3 (based on fold recognition methods). MFDp2 (http://biomine.cs.vcu.edu/servers/MFDp2/) predicts the disorder at the sequence and residue levels. AUCPred and SPOT-disorder can predict short and long IDRs. AUCpreD (http://raptorx.uchicago.edu/StructurePropertyPred/predict/) predicts disorder considering information from the sequence, evolution, and structural levels, while SPOT-disorder (https://sparks-lab.org/server/spot-disorder-single/) considers sequence-level information. These predictors assign a disorder score to each amino acid of the query sequence. The disorder probability of the Esg protein was calculated by averaging the results obtained by all of the predictors. Residues with scores >0.5 were considered to be disordered. The result was plotted using Matplotlib 2.2.3.[73]

4.2. Generation of Starting Structures for the Simulations. To obtain initial structures of the S2 region (residues 111−155: ^{111}AAAAAASVVVPTPTYPKYPWNNFHMSPYTAEFYRTINQQGHQILP^{155}) of Esg, the predictors HHPred[74] (https://toolkit.tuebingen.mpg.de/tools/hhpred), I-Tasser[75] (https://zhanglab.ccmb.med.umich.edu/I-TASSER/), Phyre2[76] (http://www.sbg.bio.ic.ac.uk/;phyre2/html/page.cgi?id=index), QUARK[77] (https://zhanglab.ccmb.med.umich.edu/QUARK/), and SPARKS-X[78] (https://sparks-lab.org/server/sparks-x/) were used. These predictors use homology modeling, fold recognition, and ab initio methods. Through this, 27 structural models were obtained as an initial ensemble, which were visualized using VMD 1.9.2;[79] the structures were prepared with CHARMM-GUI[80] for their simulation. The N- and C-termini were charged, histidine was neutral (protonated in the δ nitrogen), and glutamate and lysine were charged.

4.3. Generation of the Mutant Models. To generate the mutant models, four models of the initial ensemble that presented structural diversity were selected. The selected models had a radius of gyration (R_g) between 10 and 11 Å, and had either wandered substantially (PI) or not (HHP) from their initial structure during the 2 μs simulations of the wild-type S2. We also chose the model with the longest continuous α-helix (SP9), as well as the model with a β-sheet (SP6). The mutants E141P, E141L, E141V, E141A_T145A, R144A, T145V, and Q149P were built using CHARMM-GUI[80] in the context of the four models. The N- and C-termini were charged, histidine was neutral (protonated in the δ nitrogen), and glutamate and lysine were charged.

4.4. Molecular Dynamics Simulations. For each model, the inputs were generated using CHARMM-GUI.[80] All of the MD simulations were performed with the software NAMD 2.10[81] (http://www.ks.uiuc.edu/Research/namd/), using the CHARMM36 force field[82] (http://mackerell.umaryland.edu/charmm_ff.shtml), the generalized Born/surface area (GBSA) model of implicit solvent[83] with the default parameters suggested by NAMD, and 2 fs time steps. All simulations were carried out at a constant temperature (298 K) with a Langevin thermostat. An ε-value of 80, viscosity of 91 cps, and ionic strength of 0.15 M were used to simulate an aqueous solution environment. The SHAKE method was used to keep all bonds involving hydrogen atoms rigid. The list of neighbors was calculated with a 14 Å cutoff every 10 steps, and the noncovalent interactions had a 12 Å cutoff, with switching at 10 Å. Each model was submitted to 2000 steps of structural minimization in vacuum to eliminate clashes before the MD runs. The MD simulations were carried out for 2 μs for each model of the S2 region, and the trajectories were stored every 1 ps, yielding the 54 μs MD ensemble.

4.5. Molecular Dynamics Simulation of the Mutants. These were carried under the same conditions, except that each model was simulated during 6 μs, yielding a 24 μs ensemble for each variant. The simulations of the wild-type models were extended to 6 μs each.

4.6. α_L Sampling Characterization. The dihedral angles Φ and φ of the 54 μs MD ensemble were calculated with CHARMM38. The selected dihedral angles Φ and φ were those with at least three consecutive residues with values in the α_L region ($30° < \Phi < 100°$ and $7° < \varphi < 67°$).[42] The probability of α_L is calculated as the fraction of the ensemble that contains α_L helices.

4.7. Molecular Dynamics Simulation of $(AAQAA)_3$. A helical and extended model of $(AAQAA)_3$ was built using the Chimera software.[84] N- and C-termini were capped by acetylation and amidation, respectively, using CHARMM-GUI.[80] Each model was simulated under the same conditions as the models of the S2 region of Esg, during 16 μs.

4.8. Trajectory Analysis. The analysis of the trajectories was performed using CARMA[85] to calculate the secondary structure. CHARMM38 was used to calculate the root-mean-square deviation (RMSD) with respect to the starting structures, radius of gyration (R_g), main chain dihedral angles, contacts (carbon−carbon distances within 6 Å), and hydrogen bonds (2.4 Å distance between hydrogen and heavy atom, no angle restriction). Contacts and hydrogen bonds were calculated for atom pairs and then all pairs belonging to each

https://dx.doi.org/10.1021/acsomega.0c02051
ACS Omega 2020, 5, 18331−18341

particular residue−residue pair were added; therefore, frequencies above 100% represent more than one consistent contact or hydrogen bond present during the whole simulation ensemble. The results were plotted with Matplotlib 2.2.3.[73]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.0c02051.

Summary of IDP simulations of the past 5 years; summary of $(AAQAA)_3$ simulations; pairwise initial RMSDs; minimum pairwise RMSD between runs after 2 μs of simulation; most frequent long-distance contacts stabilizing the MoRF; and most frequent interactions between aromatic residues and prolines in the MoRF (Tables S1−S6). Contact map of the residues in the MoRF and snapshot showing one of the stabilizing long-distance contacts, for mutant E141V; snapshots of the most frequent interactions between aromatic and proline residues in the S2 region for mutant E14A_T145A; Ramachandran plot for proline residues in wild-type S2; contact map for $(AAQAA)_3$; and snapshots of the most frequent interactions in $(AAQAA)_3$ (Figures S1−S5) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

[§]**Nina Pastor** − *Laboratorio de Dinámica de Proteínas, Centro de Investigación en Dinámica Celular-IICBA, Universidad Autónoma del Estado de Morelos, 62209 Cuernavaca, México;* orcid.org/0000-0001-7755-2936; Email: nina@uaem.mx

Author

Teresa Hernández-Segura − *Laboratorio de Dinámica de Proteínas, Centro de Investigación en Dinámica Celular-IICBA and Doctorado en Ciencias CIDC-IICBA, Universidad Autónoma del Estado de Morelos, 62209 Cuernavaca, México*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.0c02051

Notes

The authors declare no competing financial interest.

[§]On sabbatical leave at Departamento de Medicina Molecular y Bioprocesos, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Av. Universidad 2001, Col. Chamilpa, 62210 Cuernavaca, Morelos, México.

■ ACKNOWLEDGMENTS

We thank CONACyT (Scholarship 559324 for T.H.-S.; grant INFR-2014-02-231509 for in-house supercomputing resources), the Clúster híbrido de SuperCómputo (CINVESTAV, Ciudad de México), and the Laboratorio Nacional de Supercómputo (LNS, Puebla) for computer time. We also thank Drs. Carlos Amero and Verónica Narváez for helpful discussions.

■ ABBREVIATIONS

MoRF:molecular recognition feature; IDP:intrinsically disordered protein; IDR:Intrinsically disordered region; Esg:Escargot; MD:molecular dynamics; MC:Monte Carlo; GBSA:generalized Born and surface area continuum solvation; ZNF:zinc finger; NTD:N-terminal domain; RMSD:root-mean-square deviation; R_g:radius of gyration

■ REFERENCES

(1) Ouyang, Y.; Zhao, L.; Zhang, Z. Characterization of the structural ensembles of p53 TAD2 by molecular dynamics simulations with different force fields. *Phys. Chem. Chem. Phys.* **2018**, *20*, 8676−8684.

(2) Shabane, P. S.; Izadi, S.; Onufriev, A. V. General Purpose Water Model Can Improve Atomistic Simulations of Intrinsically Disordered Proteins. *J. Chem. Theory Comput.* **2019**, *15*, 2620−2634.

(3) Shrestha, U. R.; Juneja, P.; Zhang, Q.; Gurumoorthy, V.; Borreguero, J. M.; Urban, V.; Cheng, X.; Pingali, S. V.; Smith, J. C.; O'Neill, H. M.; Petridis, L. Generation of the configurational ensemble of an intrinsically disordered protein from unbiased molecular dynamics simulation. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 20446−20452.

(4) Keppel, T. R.; Weis, D. D. Mapping residual structure in intrinsically disordered proteins at residue resolution using millisecond hydrogen/deuterium exchange and residue averaging. *J. Am. Soc. Mass Spectrom.* **2015**, *26*, 547−554.

(5) Gianni, S.; Dogan, J.; Jemth, P. Distinguishing induced fit from conformational selection. *Biophys. Chem.* **2014**, *189*, 33−39.

(6) Toto, A.; Malagrinò, F.; Visconti, L.; Troilo, F.; Pagano, L.; Brunori, M.; Jemth, P.; Gianni, S. Templated folding of intrinsically disordered proteins. *J. Biol. Chem.* **2020**, *295*, 6586−6593.

(7) Zhang, W.; Ganguly, D.; Chen, J. Residual structures, conformational fluctuations, and electrostatic interactions in the synergistic folding of two intrinsically disordered proteins. *PLoS Comput. Biol.* **2012**, *8*, No. e1002353.

(8) Click, T. H.; Ganguly, D.; Chen, J. Intrinsically disordered proteins in a physics-based world. *Int. J. Mol. Sci.* **2010**, *11*, 5292−5309.

(9) Pinheiro, A. S.; Marsh, J. A.; Forman-Kay, J. D.; Peti, W. Structural signature of the MYPT1-PP1 interaction. *J. Am. Chem. Soc.* **2011**, *133*, 73−80.

(10) Fuxreiter, M.; Simon, I.; Friedrich, P.; Tompa, P. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J. Mol. Biol.* **2004**, *338*, 1015−1026.

(11) Uversky, V. N. Intrinsic Disorder, Protein−Protein Interactions, and Disease. *Advances in Protein Chemistry and Structural Biology*; Academic Press Inc, 2018; pp 85−121.

(12) Frege, T.; Uversky, V. N. Intrinsically disordered proteins in the nucleus of human cells. *Biochem. Biophys. Rep.* **2015**, *1*, 33−51.

(13) Habchi, J.; Tompa, P.; Longhi, S.; Uversky, V. N. Introducing protein intrinsic disorder. *Chem. Rev.* **2014**, *114*, 6561−6588.

(14) Uversky, V. N. Dancing protein clouds: The strange biology and chaotic physics of intrinsically disordered proteins. *J. Biol. Chem.* **2016**, *291*, 6681−6688.

(15) Tompa, P.; Fersht, A. Extension of the Structure−Function Paradigm. *Structure and Function of Intrinsically Disordered Proteins*; CRC Press, 2009; pp 205−236.

(16) Staby, L.; O'Shea, C.; Willemoës, M.; Theisen, F.; Kragelund, B. B.; Skriver, K. Eukaryotic transcription factors: Paradigms of protein intrinsic disorder. *Biochem. J.* **2017**, *474*, 2509−2532.

(17) Liu, J.; Perumal, N. B.; Oldfield, C. J.; Su, E. W.; Uversky, V. N.; Dunker, A. K. Intrinsic disorder in transcription factors. *Biochemistry* **2006**, *45*, 6873−6888.

(18) Kukharenko, O.; Sawade, K.; Steuer, J.; Peter, C. Using Dimensionality Reduction to Systematically Expand Conformational Sampling of Intrinsically Disordered Peptides. *J. Chem. Theory Comput.* **2016**, *12*, 4726−4734.

(19) Best, R. B. Emerging consensus on the collapse of unfolded and intrinsically disordered proteins in water. *Curr. Opin. Struct. Biol.* **2020**, *60*, 27−38.

(20) Chong, S.-H.; Chatterjee, P.; Ham, S. Computer Simulations of Intrinsically Disordered Proteins. *Annu. Rev. Phys. Chem.* **2017**, *68*, 117−134.

(21) Das, P.; Matysiak, S.; Mittal, J. Looking at the Disordered Proteins through the Computational Microscope. *ACS Cent. Sci.* **2018**, *4*, 534−542.
(22) Bottaro, S.; Lindorff-Larsen, K.; Best, R. B. Variational optimization of an all-atom implicit solvent force field to match explicit solvent simulation data. *J. Chem. Theory Comput.* **2013**, *9*, 5641−5652.
(23) Juneja, A.; Ito, M.; Nilsson, L. Implicit solvent models and stabilizing effects of mutations and ligands on the unfolding of the amyloid β-peptide central helix. *J. Chem. Theory Comput.* **2013**, *9*, 834−846.
(24) Onufriev, A. Implicit Solvent Models in Molecular Dynamics Simulations: A Brief Overview. *Annual Reports in Computational Chemistry*; Elsevier BV, 2008; pp 125−137.
(25) Ganguly, D.; Chen, J. Atomistic details of the disordered states of KID and pKID. implications in coupled binding and folding. *J. Am. Chem. Soc.* **2009**, *131*, 5214−5223.
(26) Umezawa, K.; Ikebe, J.; Takano, M.; Nakamura, H.; Higo, J. Conformational ensembles of an intrinsically disordered protein pKID with and without a KIX domain in explicit solvent investigated by all-atom multicanonical molecular dynamics. *Biomolecules* **2012**, *2*, 104−121.
(27) Sridhar, A.; Orozco, M.; Collepardo-Guevara, R. Protein disorder-to-order transition enhances the nucleosome-binding affinity of H1. *Nucleic Acids Res.* **2020**, *48*, 5318−5331.
(28) Scholes, N. S.; Weinzierl, R. O. J. Molecular Dynamics of 'Fuzzy' Transcriptional Activator-Coactivator Interactions. *PLoS Comput. Biol.* **2016**, *12*, No. e1004935.
(29) Wang, J.; Cao, Z.; Li, S. Molecular Dynamics Simulations of Intrinsically Disordered Proteins in Human Diseases. *Curr. Comput. Aided-Drug Des.* **2009**, *5*, 280−287.
(30) Anandakrishnan, R.; Drozdetski, A.; Walker, R. C.; Onufriev, A. V. Speed of conformational change: Comparing explicit and implicit solvent molecular dynamics simulations. *Biophys. J.* **2015**, *108*, 1153−1164.
(31) Vitalis, A.; Pappu, R. V. ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions. *J. Comput. Chem.* **2009**, *30*, 673−699.
(32) Wu, K. P.; Weinstock, D. S.; Narayanan, C.; Levy, R. M.; Baum, J. Structural Reorganization of α-Synuclein at Low pH Observed by NMR and REMD Simulations. *J. Mol. Biol.* **2009**, *391*, 784−796.
(33) Mao, A. H.; Crick, S. L.; Vitalis, A.; Chicoine, C. L.; Pappu, R. V. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 8183−8188.
(34) Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J. Phys. Chem. B* **2015**, *119*, 5113−5123.
(35) Henriques, J.; Cragnell, C.; Skepö, M. Molecular Dynamics Simulations of Intrinsically Disordered Proteins: Force Field Evaluation and Comparison with Experiment. *J. Chem. Theory Comput.* **2015**, *11*, 3420−3431.
(36) Basu, S.; Biswas, P. Salt-bridge dynamics in intrinsically disordered proteins: A trade-off between electrostatic interactions and structural flexibility. *Biochim. Biophys. Acta, Proteins Proteomics* **2018**, *1866*, 624−641.
(37) Voelz, V. A.; Singh, V. R.; Wedemeyer, W. J.; Lapidus, L. J.; Pande, V. S. Unfolded-state dynamics and structure of protein L characterized by simulation and experiment. *J. Am. Chem. Soc.* **2010**, *132*, 4702−4709.
(38) Rhee, Y. M.; Sorin, E. J.; Jayachandran, G.; Lindahl, E.; Pande, V. S. Simulations of the role of water in the protein-folding mechanism. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 6456−6461.
(39) Lee, K. H.; Chen, J. Optimization of the GBMV2 implicit solvent force field for accurate simulation of protein conformational equilibria. *J. Comput. Chem.* **2017**, *38*, 1332−1341.
(40) Robustelli, P.; Piana, S.; Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, E4758−E4766.
(41) Beauchamp, K. A.; Lin, Y. S.; Das, R.; Pande, V. S. Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. *J. Chem. Theory Comput.* **2012**, *8*, 1409−1414.
(42) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; De Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **2017**, *14*, 71−73.
(43) Liu, H.; Song, D.; Zhang, Y.; Yang, S.; Luo, R.; Chen, H. F. Extensive tests and evaluation of the CHARMM36IDPSFF force field for intrinsically disordered proteins and folded proteins. *Phys. Chem. Chem. Phys.* **2019**, *21*, 21918−21931.
(44) Yang, D. J.; Chung, J. Y.; Lee, S. J.; Park, S. Y.; Pyo, J. H.; Ha, N. C.; Yoo, M. A.; Park, B. J. Slug, mammalian homologue gene of Drosophila escargot, promotes neuronal-differentiation through suppression of HEB/daughtherless. *Cell Cycle* **2010**, *9*, 2861−2874.
(45) Hemavathy, K.; Guru, S. C.; Harris, J.; Chen, J. D.; Ip, Y. T. Human Slug Is a Repressor That Localizes to Sites of Active Transcription. *Mol. Cell. Biol.* **2000**, *20*, 5087−5095.
(46) Villarejo, A.; Cortés-Cabrera, Á.; Molina-Ortíz, P.; Portillo, F.; Cano, A. Differential role of snail1 and snail2 zinc fingers in E-cadherin repression and epithelial to mesenchymal transition. *J. Biol. Chem.* **2014**, *289*, 930−941.
(47) Voog, J.; Sandall, S. L.; Hime, G. R.; Resende, L. P. F.; Loza-Coll, M.; Aslanian, A.; Yates, J. R.; Hunter, T.; Fuller, M. T.; Jones, D. L. Escargot Restricts Niche Cell to Stem Cell Conversion in the Drosophila Testis. *Cell Rep.* **2014**, *7*, 722−734.
(48) Sethi, A.; Tian, J.; Vu, D. M.; Gnanakaran, S. Identification of minimally interacting modules in an intrinsically disordered protein. *Biophys. J.* **2012**, *103*, 748−757.
(49) Navarro-Retamal, C.; Bremer, A.; Alzate-Morales, J.; Caballero, J.; Hincha, D. K.; González, W.; Thalhammer, A. Molecular dynamics simulations and CD spectroscopy reveal hydration-induced unfolding of the intrinsically disordered LEA proteins COR15A and COR15B from: *Arabidopsis thaliana*. *Phys. Chem. Chem. Phys.* **2016**, *18*, 25806−25816.
(50) Ilizaliturri-Flores, I.; Correa-Basurto, J.; Bello, M.; Rosas-Trigueros, J. L.; Zamora-López, B.; Benítez-Cardoza, C. G.; Zamorano-Carrillo, A. Mapping the intrinsically disordered properties of the flexible loop domain of Bcl-2: a molecular dynamics simulation study. *J. Mol. Model.* **2016**, *22*, No. 98.
(51) Tomasso, M. E.; Tarver, M. J.; Devarajan, D.; Whitten, S. T. Hydrodynamic Radii of Intrinsically Disordered Proteins Determined from Experimental Polyproline II Propensities. *PLoS Comput. Biol.* **2016**, *12*, No. e1004686.
(52) Guo, X.; Han, J.; Luo, R.; Chen, H. F. Conformation dynamics of the intrinsically disordered protein c-Myb with the ff99IDPs force field. *RSC Adv.* **2017**, *7*, 29713−29721.
(53) Song, D.; Luo, R.; Chen, H. F. The IDP-Specific Force Field ff14IDPSFF Improves the Conformer Sampling of Intrinsically Disordered Proteins. *J. Chem. Inf. Model.* **2017**, *57*, 1166−1178.
(54) Sherry, K. P.; Das, R. K.; Pappu, R. V.; Barrick, D. Control of transcriptional activity by design of charge patterning in the intrinsically disordered RAM region of the Notch receptor. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, E9243−E9252.
(55) Das, R. K.; Pappu, R. V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 13392−13397.
(56) Holehouse, A. S.; Das, R. K.; Ahad, J. N.; Richardson, M. O. G.; Pappu, R. V. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophys. J.* **2017**, *112*, 16−21.
(57) Carballo-Pacheco, M.; Strodel, B. Comparison of force fields for Alzheimer's A β42: A case study for intrinsically disordered proteins. *Protein Sci.* **2017**, *26*, 174−185.

https://dx.doi.org/10.1021/acsomega.0c02051
ACS Omega 2020, 5, 18331−18341

(58) Donald, J. E.; Kulp, D. W.; DeGrado, W. F. Salt bridges: Geometrically specific, designable interactions. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 898−915.
(59) Zhang, X.; Li, L.; Li, N.; Shu, X.; Zhou, L.; Lü, S.; Chen, S.; Mao, D.; Long, M. Salt bridge interactions within the β2 integrin α7 helix mediate force-induced binding and shear resistance ability. *FEBS J.* **2018**, *285*, 261−274.
(60) Basu, S.; Mukharjee, D. Salt-bridge networks within globular and disordered proteins: characterizing trends for designable interactions. *J. Mol. Model.* **2017**, *23*, No. 206.
(61) Huerta-Viga, A.; Amirjalayer, S.; Domingos, S. R.; Meuzelaar, H.; Rupenyan, A.; Woutersen, S. The structure of salt bridges between Arg+ and Glu- in peptides investigated with 2D-IR spectroscopy: Evidence for two distinct hydrogen-bond geometries. *J. Chem. Phys.* **2015**, *142*, No. 212444.
(62) Vijayakumar, M.; Qian, H.; Zhou, H. X. Hydrogen bonds between short polar side chains and peptide backbone: Prevalence in proteins and effects on helix-forming propensities. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 497−507.
(63) Wang, J.; Feng, J. A. Exploring the sequence patterns in the α-helices of proteins. *Protein Eng.* **2003**, *16*, 799−807.
(64) Fuse, N.; Hirose, S.; Hayashi, S. Diploidy of Drosophila imaginal cells is maintained by a transcriptional repressor encoded by escargot. *Genes Dev.* **1994**, *8*, 2270−2281.
(65) Mateos, B.; Conrad-Billroth, C.; Schiavina, M.; Beier, A.; Kontaxis, G.; Konrat, R.; Felli, I. C.; Pierattelli, R. The Ambivalent Role of Proline Residues in an Intrinsically Disordered Protein: From Disorder Promoters to Compaction Facilitators. *J. Mol. Biol.* **2019**, *432*, 3093−3111.
(66) Huang, J.; Mackerell, A. D. Induction of peptide bond dipoles drives cooperative helix formation in the (AAQAA)3 peptide. *Biophys. J.* **2014**, *107*, 991−997.
(67) Bateman, A. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **2019**, *47*, D506−D515.
(68) Kozlowski, L. P.; Bujnicki, J. M. MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. *BMC Bioinformatics* **2012**, *13*, No. 111.
(69) Mizianty, M. J.; Peng, Z.; Kurgan, L. MFDp2: Accurate predictor of disorder in proteins by fusion of disorder probabilities, content and profiles. *Intrinsically Disord. Proteins* **2013**, *1*, No. e24428.
(70) Wang, S.; Ma, J.; Xu, J. AUCpreD: Proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields. *Bioinformatics* **2016**, i672−i679.
(71) Hanson, J.; Yang, Y.; Paliwal, K.; Zhou, Y. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics* **2017**, *33*, 685−692.
(72) Nielsen, J. T.; Mulder, F. A. A. Quality and bias of protein disorder predictors. *Sci. Rep.* **2019**, *9*, No. 5137.
(73) Caswell, T. A.; Droettboom, M.; Hunter, J.; Firing, E.; Lee, A.; Nielsen, J. H.; Andrade, E. S.; de Stansby, D.; Varoquaux, N.; Klymak, J.; Root, B.; Elson, P.; Dale, D.; Lee, J.-J.; May, R.; Seppänen, J. K.; McDougall, D.; Straw, A.; Hoffmann, T.; Hobson, P.; cgohlke; Yu, T. S.; Ma, E.; Vicent, A. F.; Silvester, S.; Moad, C.; Katins, J.; Kniazev, N.; Ariza, F.; Würtz, P. *Matplotlib/Matplotlib*, version 2.2.3; Zenodo, 2018. http://doi.org/10.5281/zenodo.1343133.
(74) Söding, J.; Biegert, A.; Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **2005**, *33*, W244−W248.
(75) Yang, J.; Zhang, Y. I-TASSER server: New development for protein structure and function predictions. *Nucleic Acids Res.* **2015**, *43*, W174−W181.
(76) Kelley, L. A.; Mezulis, S.; Yates, C. M.; Wass, M. N.; Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **2015**, *10*, 845−858.
(77) Xu, D.; Zhang, Y. Toward optimal fragment generations for ab initio protein structure assembly. *Proteins: Struct., Funct., Bioinf.* **2013**, *81*, 229−239.
(78) Yang, Y.; Faraggi, E.; Zhao, H.; Zhou, Y. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics* **2011**, *27*, 2076−2082.
(79) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular. *Dynamics* **1996**, *14*, 33−38.
(80) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: A web-based graphical user interface for CHARMM. *J. Comput. Chem.* **2008**, *29*, 1859−1865.
(81) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781−1802.
(82) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone φ, ψ and side chain χ1 and χ2 Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8*, 3257−3273.
(83) Tanner, D. E.; Chan, K. Y.; Phillips, J. C.; Schulten, K. Parallel generalized born implicit solvent calculations with NAMD. *J. Chem. Theory Comput.* **2011**, *7*, 3635−3642.
(84) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.
(85) Glykos, N. M. Software news and updates carma: A molecular dynamics analysis program. *J. Comput. Chem.* **2006**, *27*, 1765−1768.

https://dx.doi.org/10.1021/acsomega.0c02051
ACS Omega 2020, 5, 18331−18341

Supporting Information:

Identification of an α-MoRF in the intrinsically disordered region of the Escargot transcription factor.

Teresa Hernandez-Segura[#] and Nina Pastor[*,&]

Laboratorio de Dinámica de Proteínas, Centro de Investigación en Dinámica Celular-IICBA, Universidad Autónoma del Estado de Morelos, Av. Universidad 1001, Chamilpa, 62209 Cuernavaca, México.

[#]Doctorado en Ciencias, CIDC-IICBA, Universidad Autónoma del Estado de Morelos, Cuernavaca, Morelos, México

[&]on sabbatical leave at Departamento de Medicina Molecular y Bioprocesos, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Av. Universidad 2001, Col. Chamilpa, 62210 Cuernavaca, Morelos, México

[*]corresponding author (nina@uaem.mx)

Table S1: MD simulation protocols of IDPs in the last five years

System	Solvent	Force field	Cumulative total time	Reference
Bcl-2 (239 residues)	Explicit (TIP3P)	CHARMM27	25 ns	1
rFLD region (residues 60-77)			1.0 μs	
Human NCOR1 CoRNR box 3 (residues 2258 – 2277)	Explicit (TIP3P)	CHARMM22*	8 μs	2
COR15A (89 residues)	Explicit (Glycerol-TIP4P)	OPLS-AA	300 ns for each system	3
COR15B (90 residues)				
α-synuclein (140 residues)	Implicit (GBSA)	CHARMM27	10 ns for each system	4
Translocation domain of Colicin N (90 residues)				
K18 domain of Tau protein (130 residues)			30ns	
α-synuclein (residues 42 - 63)	Explicit (SPC water)	GROMOS 54A7	14 μs	5
C-Myb (residues 291-315)	Explicit (TIP3P, TIP4P-EW, TIP5P)	ff99IDPs ff99SBildn	1.9 μs (TIP3P) 950 ns (TIP4P-EW) 950 ns (TIP5P)	6
			1 μs (TIP3P) 500 ns (TIP4P-Ew) 500 ns (TIP5P)	
Aβ□□40 (30 residues)	Explicit (TIP3P, TIP3Pm)	CHARMM36	3.2 μs (TIP3P) 3.2 μs (TIP3Pm)	7
		CHARMM22	3.2 μs (TIP3P)	
		CHARMM22*	3.2 μs (TIP3Pm)	
		OPLS-AA	3.2 μs (TIP3Pm)	
Unfolded peptide: PepG (9 residues)	Explicit (TIP4P)	ff03w	14.04 μs	8
Unfolded peptide: PepW (9 residues)			4.04 μs	
p53 CTD domain (residues 375 – 388)			9.64 μs	
Aβ42 (42 residues)	Explicit (TIP4P-Ew)	ff99SB*ILDN ff99SBILDN-NMR ff99SB CHARMM22*	6.4 μs for each force field	9
	(TIP3P)	OPLS ff99SB		
RAM disordered region of the Notch receptor (WT and 13 variants) (residues 1-100)	Implicit (ABSINTH)	OPLS-AA	9.20 x 10^8 steps MMC for each system	10
γ-tubulin Carboxyl-terminus (WT and 3 variants) (35 residues)	Implicit (ABSINTH)	OPLS-AA	1.20x10^8 steps MMC for each system	11
9 short peptides (9 residues)	Explicit (TIP3P)	ff14SB ff14IDPSFF	10 μs for each for each system	12
HIV-1 Rev protein (23 residues)			1 μs	
Aβ40 (40 residues)	Explicit	ff99SBws	743.760 ns (Aβ40)	13

	(TIP4P)		740.715 ns (Aβ42)	
Aβ42 (42 residues)		ff03ws	750 ns for each system	
p53 TAD2 (residues 41-62)	Explicit (TIP3P)	ff03 CHARMM27 OPLS-AA/L ff99SB-ILDN CHARMM36m	3 μs for each force field	[14]
Syt-1IDR (residues 80-141)	Implicit (GBSA)	CHARMM36	3.180 μs	[15]
Core region of the IDR (residues 97-130)			4.18 μs	
$A\beta_{1-40}$ (WT and one variant)	Implicit (GBSA)	CHARMM36	1.098 μs for each system	[16]
SH4UD (95 residues)	Explicit (TIP3P)	ff03ws	10.20 μs	[17]
Protein IN (55 residues) $A\beta_{1-42}$ (42 residues) H4 histone tail (26 residues)	Explicit (TIP3P, OPC)	ff99SB	for each water model: 15 μs (Protein IN) 10 μs ($A\beta_{1-42}$ and H4 histone tail)	[18]
	Implicit (GB Neck)		1 μs for each system	
N_{tail} α-MoRE (residues 484 to 504)	Explicit (TIP4P-D)	ff99SB	100 μs	[19]
Human H1.0 NTD subtype (10 residues) Human H1.1 NTD subtype (21 residues) Human H1.2 NTD subtype (18 residues)	Explicit (TIP3P, Modeled TIP3)	ff99SB	14 μs for H1.0 24 μs for H1.1 18 μs for H1.2	[20]
		CHARMM36m		
4E-BP2 (WT and two variants) (residues 18 to 62)	Explicit (TIP3P)	CHARMM36m	19.20 μs for each system and force field	[21]
		ff99SB-ILDN		

ABBREVIATIONS:

rFLD=region of flexible loop domain (FLD)
MMC =Metropolis Monte Carlo
Bcl-2 = B-cell lymphoma-2 protein
COR15A= Lea protein COR15A
COR15B =Lea protein COR15B
NCOR1 CoRNR box 3 = Nuclear receptor corepressor 1 (NCOR1) CoRNR box 3 motif
Protein IN= N-terminal zinc-binding domain of HIV-1
SH4UD = N-terminal of c-Src Kinase
N_{tail} α-MoRE = α-helical Molecular Recognition Element (α-MoRE) of the intrinsically disordered C-terminal domain of the measles virus nucleoprotein (N_{TAIL})
4E-BP2= 4E-binding protein 2

Table S2: MD simulation protocols of $(AAQAA)_3$ and their % helicity

System	Solvent	Force field	Cumulative time total	% helicity	Reference
$(AAQAA)_3$	Implicit (GBSW)	CHARMM22/CMAP	320 ns	~65%	[22]
	Explicit (TIP3P)			~90	
	Explicit (TIP3P)	ff03	960 ns for each force field	93.9%	[23]
		ff99SB		26.9%	
		ff03*		45.9%	
		ff99SB*		48.5%	
	Explicit (TIP3P)	CHARMM36	4.8 μs	~32%	[24]
	Explicit (TIP3P)	CHARMM36	4.8 μs	~44%	[25]
	Implicit (EEF1-C19, EEF1-SB FACTS, SCPISM)	CHARMM36/EEF1-C9	100 ns for each force field	~12%	
		CHARMM36/EEF1-SB		~30%	
		CHARMM36/FACTS		~90%	
		CHARMM36/SCPISM		~85%	
	Explicit (TIP3P)	CHARMM36m (C36m)	16 μs for each force field	17%	[26]
		CHARMM36 (C36)		13%	
	Implicit (GBMV2)	CHARMM36	320 ns	42-47%	[27]
	Explicit (TIP3P/TIP4P)	C22*/TIP3P	20 μs for each force field	~30%	[28]
		C36m/TIP3P a99SB-ILDN/TIP3P a03ws/TIP4P-D a99SB-ILDN/TIP4P-D a99SB/TIP4P-Ew		5-12%	
	Explicit (Modified TIP3P)	C36IDPSFF	5 μs	~10%	[29]

ABBREVIATIONS:

ff99SB* and ff03* are force fields with corrections in φ

Table S3: Comparison of the pairwise RMSD of the initial structures of the S2 region. The red and blue squares indicate the lowest and highest pairwise RMSD, respectively.

Matrix of pairwise RMSD (in Å) between Cα of the initial structures of the S2 region.																										
	HHP	IT1	IT2	IT3	IT4	IT5	PI	Q1	Q10	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	SP1	SP10	SP2	SP3	SP4	SP5	SP6	SP7	SP8	SP9
HHP		9.5	8.48	9.6	8.6	8.5	9.6	9.1	7.5	5.8	9.3	7.5	7.76	5.9	7.0	7.8	8.7	8.4	9.8	3.0	10.4	3.8	7.8	9.8	10.9	8.2	11.8
IT1	9.5		3.70	3.0	4.6	5.5	9.0	9.2	6.3	8.3	8.6	6.8	8.11	8.3	6.3	5.2	5.4	7.4	9.1	9.9	10.1	9.7	7.1	8.7	10.9	7.2	12.3
IT2	8.5	3.7		3.7	3.9	6.1	8.8	9.1	5.7	7.0	8.6	6.4	7.6	7.2	5.4	4.8	6.0	6.8	8.7	8.6	10.3	8.1	7.4	8.8	11.0	6.8	12.2
IT3	9.6	3.0	3.7		4.6	5.8	8.6	9.1	6.2	8.3	8.2	7.4	7.8	8.7	6.8	4.9	6.3	7.1	8.9	9.9	10.6	9.6	7.3	8.5	10.9	7.8	12.1
IT4	8.6	4.6	3.9	4.6		5.5	8.4	9.6	6.5	7.0	8.5	85.8	8.0	6.9	5.4	4.7	5.3	7.4	9.2	8.6	11.1	8.4	7.7	8.2	10.9	6.7	12.6
IT5	8.5	5.5	6.1	5.8	5.5		8.1	9.3	7.8	8.6	8.3	6.5	8.7	7.8	5.6	6.7	4.8	7.7	8.4	8.9	9.8	8.8	6.4	9.1	10.4	8.2	11.6
PI	9.6	9.0	8.8	8.6	8.4	8.1		9.4	9.4	10.2	10.6	9.1	10.4	9.6	8.7	7.5	8.1	8.5	8.1	10.3	11.0	9.7	8.0	8.2	7.8	10.2	11.9
Q1	9.1	9.2	9.1	9.1	9.6	9.3	9.4		8.2	8.0	4.1	9.0	6.1	9.1	8.7	8.6	8.7	8.5	9.1	9.6	8.6	8.9	8.4	7.8	7.2	8.8	13.0
Q10	7.5	6.3	5.7	6.2	6.5	7.8	9.4	8.2		5.1	7.4	7.5	6.0	6.9	6.8	6.3	7.5	7.4	9.6	7.2	10.5	6.8	8.6	9.8	9.8	8.3	10.4
Q2	5.8	8.3	7.0	8.3	7.0	8.6	10.2	8.0	5.1		7.2	7.0	4.6	4.1	6.5	6.2	8.5	7.8	9.6	5.4	9.2	4.9	8.7	10.2	9.6	7.8	11.3
Q3	9.3	8.6	8.6	8.2	8.5	8.3	10.6	4.1	7.4	7.2		8.6	5.5	8.4	8.5	7.9	8.6	9.5	10.0	9.4	8.9	8.4	8.4	8.2	8.0	9.1	12.5
Q4	7.5	6.8	6.4	7.4	5.8	6.5	9.1	9.0	7.5	7.0	8.6		8.8	4.8	4.6	6.5	4.1	7.9	8.7	8.4	10.7	8.1	6.0	8.1	10.9	6.3	12.5
Q5	7.8	8.1	7.6	7.8	8.0	8.7	10.4	6.1	6.0	4.6	5.5	8.8		6.2	8.1	6.8	8.9	7.7	8.7	7.3	7.8	6.6	9.5	9.8	8.1	8.8	12.2
Q6	5.9	8.3	7.2	8.7	6.9	7.8	9.6	9.1	6.9	4.1	8.4	4.8	6.2		5.3	6.3	6.9	7.5	9.3	6.5	9.4	6.0	7.3	9.5	10.7	7.0	12.1
Q7	7.0	6.3	5.4	6.8	5.4	5.6	8.7	8.7	6.8	6.5	8.5	4.6	8.1	5.3		6.3	5.3	7.0	8.4	7.1	10.4	7.1	6.3	8.4	9.9	7.9	11.9
Q8	7.8	5.2	4.8	4.9	4.7	6.7	7.5	8.6	6.3	6.2	7.9	6.5	6.8	6.3	6.3		5.8	8.6	10.6	8.2	11.2	7.6	7.9	7.3	9.1	6.1	11.9
Q9	8.7	5.4	6.0	6.3	5.3	4.8	8.1	8.7	7.5	8.5	8.6	4.1	8.9	6.9	5.3	5.8		8.3	9.1	9.6	10.6	9.4	5.9	7.9	9.6	6.7	12.7
SP1	8.4	7.4	6.8	7.1	7.4	7.7	8.5	8.5	7.4	7.8	9.5	7.9	7.7	7.5	7.0	8.6	8.3		6.1	8.6	8.0	8.5	6.5	11.1	9.9	9.3	11.8
SP10	9.8	9.1	8.7	8.9	9.2	8.4	8.1	9.1	9.6	9.6	10.0	8.7	8.7	9.3	8.4	10.6	9.1	6.1		9.8	6.7	9.8	7.8	10.4	8.5	10.5	13.0
SP2	3.0	9.9	8.6	9.9	8.6	8.9	10.3	9.6	7.2	5.4	9.4	8.4	7.3	6.5	7.1	8.2	9.6	8.6	9.8		10.1	3.3	8.6	10.6	11.0	8.8	12.0
SP3	10.4	10.1	10.3	10.6	11.1	9.8	11.0	8.6	10.5	9.2	8.9	10.7	7.8	9.4	10.4	11.2	10.6	8.0	6.7	10.1		10.1	9.5	12.9	10.4	11.0	12.4
SP4	3.8	9.7	8.1	9.6	8.4	8.8	9.7	8.9	6.8	4.9	8.4	8.1	6.6	6.0	7.1	7.6	9.4	8.5	9.8	3.3	10.1		8.4	10.2	10.2	8.0	12.3
SP5	7.8	7.1	7.4	7.3	7.7	6.4	8.0	8.4	8.6	8.7	8.4	6.0	9.5	7.3	6.3	7.9	5.9	6.5	7.8	8.6	9.5	8.4		8.4	10.3	8.2	11.3
SP6	9.8	8.7	8.8	8.5	8.2	9.1	8.2	7.8	9.8	10.2	8.2	8.1	9.8	9.5	8.4	7.3	7.9	11.1	10.4	10.6	12.9	10.2	8.4		7.4	8.8	13.9

SP7	10.9	10.9	11.0	10.9	10.9	10.4	7.8	7.2	9.8	9.6	8.0	10.9	8.1	10.7	9.9	9.1	9.6	9.9	8.5	11.0	10.4	10.2	10.3	7.4		10.9	12.6
SP8	8.2	7.2	6.8	7.8	6.7	8.2	10.2	8.8	8.3	7.8	9.1	6.1	8.8	7.0	7.9	6.1	6.7	9.3	10.5	8.8	11.0	8.0	8.2	8.8	10.9		14.1
SP9	11.8	12.3	12.2	12.1	12.6	11.6	11.9	13.0	10.4	11.3	12.5	12.5	12.2	12.1	11.9	12.0	12.7	11.8	13.0	12.0	12.4	12.3	11.3	13.9	12.6	14.1	

Table S4: Minimum pairwise RMSD during 2μs of simulation of each model of the S2 region ensemble, compared to the structures generated by the simulations of the other models. The orange squares indicate the lowest pairwise RMSD between each pair of simulations. The green and yellow squares indicate the smallest and largest of the minimum of pairwise RMSDs.

Matrix of pairwise RMSD (in Å) between Cα during 2μs of simulation of each structure of the S2 region.																										
	HHP	IT1	IT2	IT3	IT4	IT5	PI	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	SP1	SP2	SP3	SP4	SP5	SP6	SP7	SP8	SP9
HHP		5.19	5.96	7.23	6.24	6.78	6.36	7.85	4.36	5.46	4.94	5.36	5.18	4.39	5.94	4.35	5.87	6.22	3.77	5.33	4.20	5.18	6.34	6.20	5.32	5.60
IT1	5.19		3.57	3.69	4.10	5.56	6.01	6.20	5.99	6.39	4.41	5.61	4.47	3.76	3.84	3.68	6.56	5.78	6.23	5.58	5.35	3.63	5.69	6.92	5.01	5.12
IT2	5.96	3.57		3.36	3.68	4.18	5.47	6.69	5.64	6.38	3.89	5.46	3.80	3.74	2.98	4.61	5.09	5.58	5.69	5.58	5.40	3.59	4.65	6.10	3.21	3.00
IT3	7.23	3.69	3.36		4.38	5.09	5.80	7.04	5.36	6.01	4.50	5.54	3.67	3.86	3.11	4.94	5.97	5.19	6.44	5.23	5.71	4.19	5.11	6.39	4.35	4.03
IT4	6.24	4.10	3.68	4.38		2.65	4.66	5.96	5.79	6.64	5.12	5.88	4.76	4.66	3.31	5.71	4.27	5.85	6.52	6.09	5.43	4.14	4.95	4.42	4.89	4.75
IT5	6.78	5.56	4.18	5.09	2.65		3.67	4.82	5.76	5.44	4.44	5.21	5.05	4.30	4.72	5.49	4.74	5.46	5.40	4.90	5.73	4.99	4.86	4.98	4.95	4.43
PI	6.36	6.01	5.47	5.80	4.66	3.67		4.23	4.86	3.68	5.61	3.83	5.44	5.53	5.66	7.07	3.62	3.87	4.06	3.90	5.20	6.64	5.80	3.75	6.30	6.08
Q1	7.85	6.20	6.69	7.04	5.96	4.82	4.23		5.13	3.65	6.87	6.03	6.54	6.10	6.18	6.92	3.81	5.03	4.80	3.43	6.04	5.15	6.46	5.60	6.58	6.96
Q2	4.36	5.99	5.64	5.36	5.79	5.76	4.86	5.13		3.30	5.03	4.22	5.07	4.41	5.97	6.11	5.45	5.10	3.59	4.11	3.52	6.29	6.15	5.35	6.10	5.35
Q3	5.46	6.39	6.38	6.01	6.64	5.44	3.68	3.65	3.30		5.75	3.21	5.94	5.22	5.77	7.00	3.82	5.15	3.88	3.24	4.25	6.78	6.64	5.66	6.28	5.83
Q4	4.94	4.41	3.89	4.50	5.12	4.44	5.61	6.87	5.03	5.75		4.03	2.70	2.86	4.36	3.70	6.79	5.36	6.35	5.86	3.52	5.00	5.77	6.22	3.60	3.59
Q5	5.36	5.61	5.46	5.54	5.88	5.21	3.83	6.03	4.22	3.21	4.03		3.57	3.48	5.46	5.51	6.53	3.89	4.68	4.06	3.98	6.33	6.05	5.50	5.48	4.58
Q6	5.18	4.47	3.80	3.67	4.76	5.05	5.44	6.54	5.07	5.94	2.70	3.57		3.01	3.83	4.00	5.14	4.86	5.37	5.14	4.07	3.75	5.64	5.20	3.88	3.80
Q7	4.39	3.76	3.74	3.86	4.66	4.30	5.53	6.10	4.41	5.22	2.86	3.48	3.01		4.16	4.50	5.69	5.65	6.00	4.85	3.82	4.46	5.24	5.70	4.18	3.97
Q8	5.94	3.84	2.98	3.11	3.31	4.72	5.66	6.18	5.97	5.77	4.36	5.46	3.83	4.16		4.99	5.37	5.68	6.02	6.05	5.23	3.67	5.17	5.72	3.30	3.56
Q9	4.35	3.68	4.61	4.94	5.71	5.49	7.07	6.92	6.11	7.00	3.70	5.51	4.00	4.50	4.99		6.77	7.43	7.39	6.88	4.49	3.93	5.98	6.78	3.80	4.26
Q10	5.87	6.56	5.09	5.97	4.27	4.74	3.62	3.81	5.45	3.82	6.79	6.53	5.14	5.69	5.37	6.77		4.71	4.82	4.60	6.28	4.58	6.50	4.84	6.06	5.66
SP1	6.22	5.78	5.58	5.19	5.85	5.46	3.87	5.03	5.10	5.15	5.36	3.89	4.86	5.65	5.68	7.43	4.71		4.77	3.82	5.60	6.55	6.28	6.02	6.42	6.00
SP2	3.77	6.23	5.69	6.44	6.52	5.40	4.06	4.80	3.59	3.88	6.35	4.68	5.37	6.00	6.02	7.39	4.82	4.77		2.87	3.43	6.14	6.64	5.68	6.02	4.86
SP3	5.33	5.58	5.58	5.23	6.09	4.90	3.90	3.43	4.11	3.24	5.86	4.06	5.14	4.85	6.05	6.88	4.60	3.82	2.87		4.35	5.74	5.78	5.35	5.44	5.03
SP4	4.20	5.35	5.40	5.71	5.43	5.73	5.20	6.04	3.52	4.25	3.52	3.98	4.07	3.82	5.23	4.49	6.28	5.60	3.43	4.35		5.73	6.29	5.98	3.81	4.47
SP5	5.18	3.63	3.59	4.19	4.14	4.99	6.64	5.15	6.29	6.78	5.00	6.33	3.75	4.46	3.67	3.93	4.58	6.55	6.14	5.74	5.73		4.41	6.06	4.01	3.89
SP6	6.34	5.69	4.65	5.11	4.95	4.86	5.80	6.46	6.15	6.64	5.77	6.05	5.64	5.24	5.17	5.98	6.50	6.28	6.64	5.78	6.29	4.41		5.65	5.86	5.47
SP7	6.20	6.92	6.10	6.39	4.42	4.98	3.75	5.60	5.35	5.66	6.22	5.50	5.20	5.70	5.72	6.78	4.84	6.02	5.68	5.35	5.98	6.06	5.65		5.39	6.04

SP8	5.32	5.01	3.21	4.35	4.89	4.95	6.30	6.58	6.10	6.28	3.60	5.48	3.88	4.18	3.30	3.80	6.06	6.42	6.02	5.44	3.81	4.01	5.86	5.39		2.55
SP9	5.60	5.12	3.00	4.03	4.75	4.43	6.08	6.96	5.35	5.83	3.59	4.58	3.80	3.97	3.56	4.26	5.66	6.00	4.86	5.03	4.47	3.89	5.47	6.04	2.55	
SP10	5.63	4.89	4.35	4.35	5.34	5.09	4.49	6.30	4.97	5.65	5.69	5.39	4.99	4.81	4.57	6.32	5.63	4.26	4.67	4.34	6.22	5.19	5.87	5.84	4.65	4.47

Table S5. Most frequent long-range contacts (n → n + 4 or more) between the α-MoRF region (from residues 134 to 152) and the rest of the S2 region, for each 24 μs ensemble. The frequency was calculated for all pairs of carbon atoms within 6 Å, and then added for all carbon-carbon pairs of the interacting residues.

Wild-type		E141L		E141P		E141V		E141A-T145A		R144A		T145V		Q149P	
Pair	%	Pair	%	Pair	%	Pair	%	Pair	%	Pair	%	Pair	%	Pair	%
P123-Y138	94	F133-Y143	676	Y128-Y143	1217	F133-Y143	807	Y128-F142	2148	W130-F142	741	F133-Y143	780	W130-Y138	878
V119-F142	88	W130-F142	598	W130-H134	577	W130-Y138	533	P129-Y138	2042	W130-Y138	726	W130-F142	653	W130-P137	665
P121-Y143	69	V120-F142	548	V119-Y138	473	V120-F142	506	Y128-Y138	1998	Y125-Y138	563	N132-F142	517	W130-F142	560
V118-F142	60	W130-H134	498	Y125-Y138	467	W130-F142	418	Y125-I146	1528	W130-Y143	481	F133-F142	443	W130-R144	553
P121-F142	48	P121-Y138	479	F133-P141	465	W130-Y143	362	Y128-H134	1204	Y128-Y143	417	P123-F142	390	W130-M135	503

Note: Numbers exceeding 100% represent more than one carbon-carbon contact between the pair of residues. For example, Y128 and F142 engaged in a little over 21 carbon-carbon contacts in the simulations for the double mutant E141A-T145A.

Table S6. Most frequent contacts (i±1 and i-2) between aromatic residues and Prolines in the S2 region for each 24 μs ensemble. The frequency was calculated for all pairs of carbon atoms within 6 Å, and then added for all carbon-carbon pairs of the interacting residues.

System	% Frequency				
	Y125-P126	Y128-P129	P129-W130	P137-Y138	P141-Y142
Wild-type	1763	1549	1543	2238	-
E141L	1259	1507	1153	1829	-
E141P	1717	1801	763	2572	1895
E141V	1361	1508	1012	1692	-
E141A_T145A	2562	2250	1949	3187	-
Q149P	1576	1640	1047	1758	-
R144A	1462	1232	1165	1788	-
T145V	1432	1237	1248	2193	-

Note: Numbers exceeding 100% represent more than one carbon-carbon contact between the pair of residues.

Figure S1. A) Heat map representing the residue-residue interactions within the α-MoRF region (from residues 134 to 152) of the E141V mutant S2. B) Conformation that represents a frequent long-range interaction of the α-MoRF region with the rest of S2. The protein backbone is depicted as a gray ribbon and the interacting residues as van der Waals spheres in CPK colors. The mutated residue is shown in red balls and sticks.

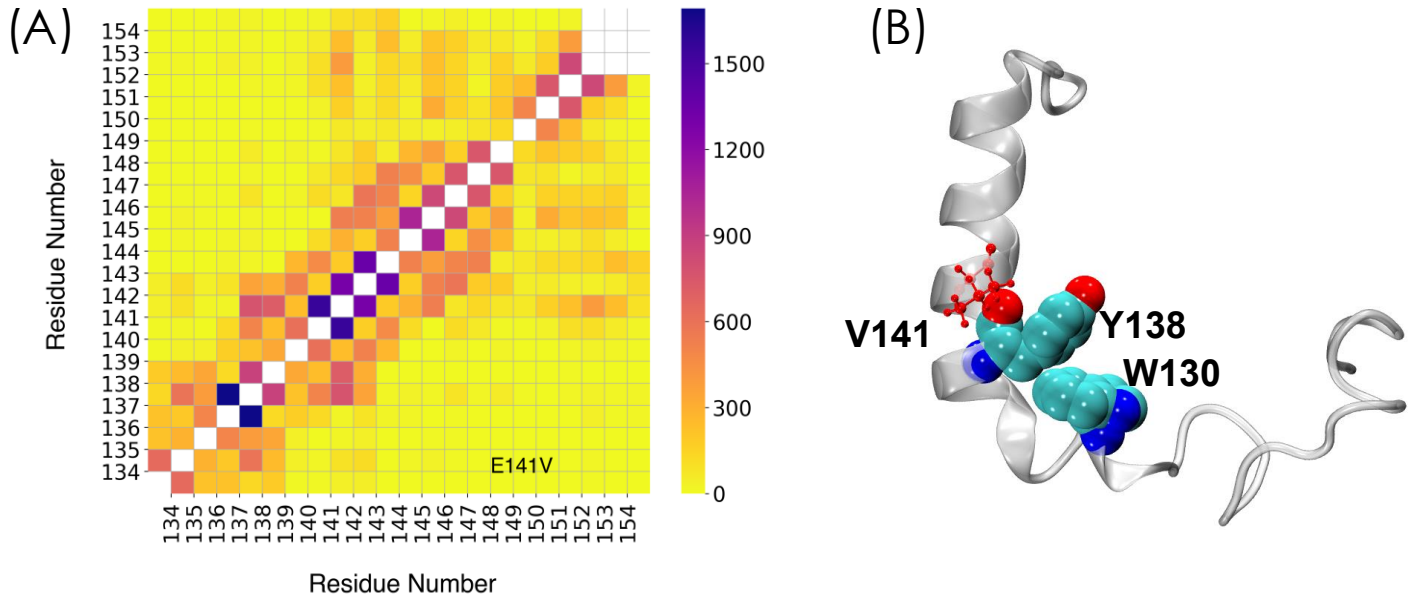


Figure S2. Snapshots showing frequent interactions between aromatic residues and Prolines in the S2 region of the E141A_T145A mutant. The protein backbone is depicted as a gray ribbon and the interacting residues as van der Waals spheres: Tyrosine in green, Prolines in purple and Tryptophan in iceblue. The mutated residues are shown in red balls and sticks.

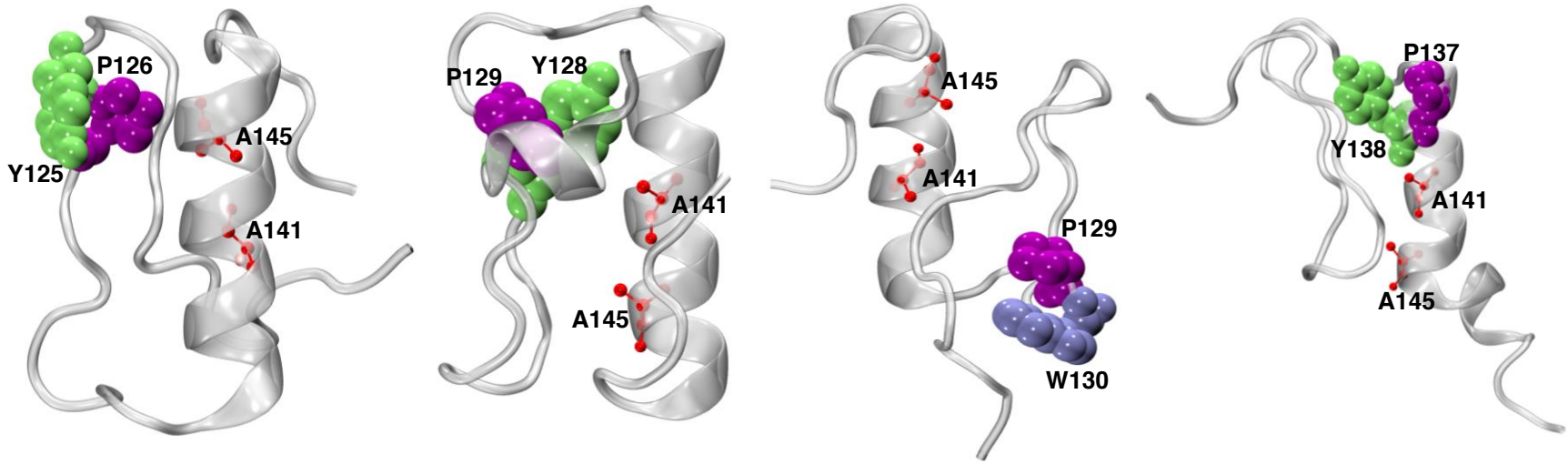


Figure S3. Ramachandran plot for Proline residues (P121, P123, P126, P129, and P137) in the 54 μs ensemble for wild-type S2.

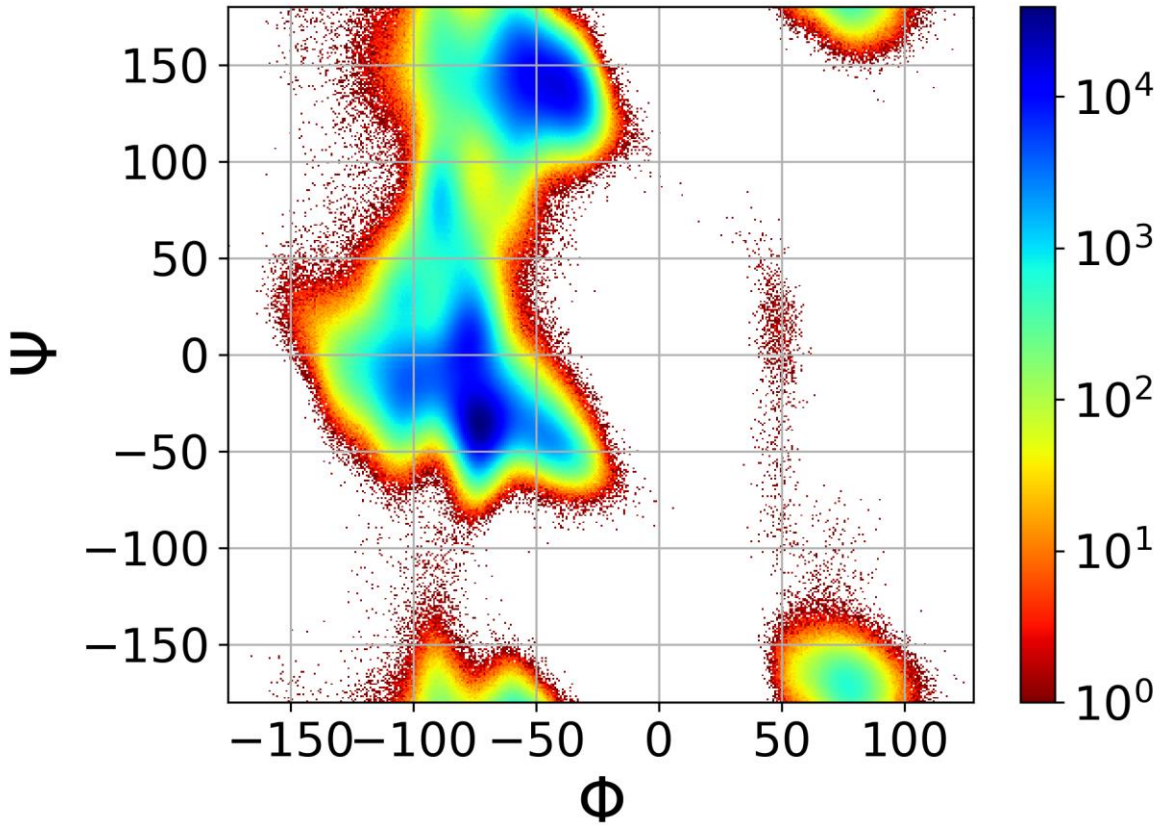


Figure S4: Contact map (carbon – carbon distances within 6Å) for the 32 μs simulation of $(AAQAA)_3$. Immediate neighbors are not considered.

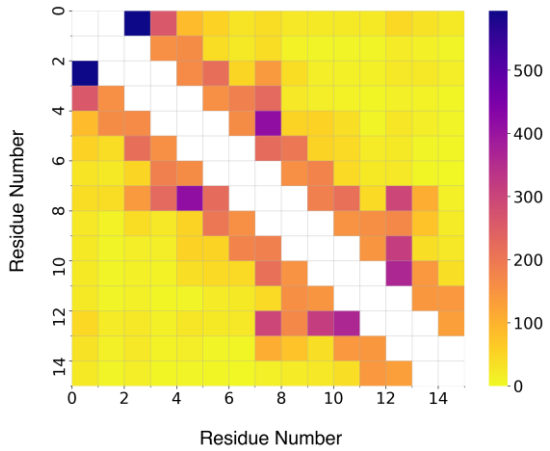
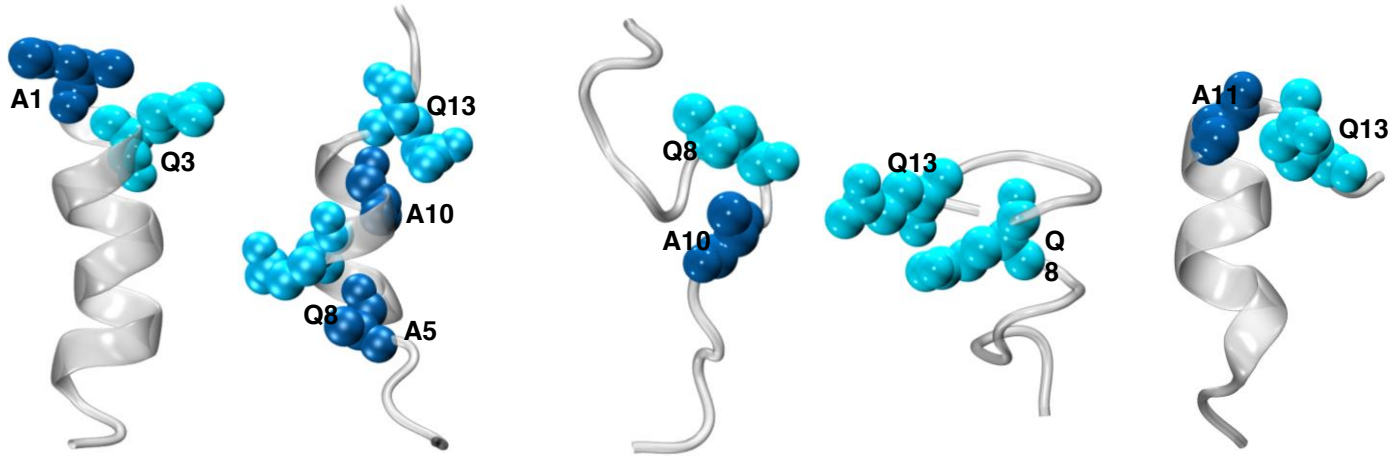


Figure S5: Most frequent inter-residue interactions in the contact map (Figure S4) for $(AAQAA)_3$. The protein backbone is shown as a gray transparent ribbon, and Alanine and Glutamine as van der Waals spheres (blue and cyan, respectively).



REFERENCES

(1) Ilizaliturri-Flores I, Correa-Basurto J, Bello M, Rosas-Trigueros JL, Zamora-López B, Benítez-Cardoza CG *et al.* Mapping the intrinsically disordered properties of the flexible loop domain of Bcl-2: a molecular dynamics simulation study. *J Mol Model* 2016; 22: 98.

(2) Cino EA, Choy WY, Karttunen M. Characterization of the Free State Ensemble of the CoRNR Box Motif by Molecular Dynamics Simulations. *J Phys Chem B* 2016; 120: 1060–1068.

(3) Navarro-Retamal C, Bremer A, Alzate-Morales J, Caballero J, Hincha DK, González W *et al.* Molecular dynamics simulations and CD spectroscopy reveal hydration-induced unfolding of the intrinsically disordered LEA proteins COR15A and COR15B from: Arabidopsis thaliana. *Phys Chem Chem Phys* 2016; 18: 25806–25816.

(4) Gong H, Zhang S, Wang J, Gong H, Zeng J. Constructing structure ensembles of intrinsically disordered proteins from chemical shift data. In: *Journal of Computational Biology*. Mary Ann Liebert Inc., 2016, pp 300–310.

(5) Kukharenko O, Sawade K, Steuer J, Peter C. Using Dimensionality Reduction to Systematically Expand Conformational Sampling of Intrinsically Disordered Peptides. *J Chem Theory Comput* 2016; 12: 4726–4734.

(6) Guo X, Han J, Luo R, Chen HF. Conformation dynamics of the intrinsically disordered protein c-Myb with the ff99IDPs force field. *RSC Adv* 2017; 7: 29713–29721.

(7) Siwy CM, Lockhart C, Klimov DK. Is the Conformational Ensemble of Alzheimer's Aβ10-40 Peptide Force Field Dependent? *PLoS Comput Biol* 2017; 13: e1005314.

(8) Han M, Xu J, Ren Y. Sampling conformational space of intrinsically disordered proteins in explicit solvent: Comparison between well-tempered ensemble approach and solute tempering method. *J Mol Graph Model* 2017; 72: 136–147.

(9) Carballo-Pacheco M, Strodel B. Comparison of force fields for Alzheimer's A β42: A case study for intrinsically disordered proteins. *Protein Sci* 2017; 26: 174–185.

(10) Sherry KP, Das RK, Pappu R V., Barrick D. Control of transcriptional activity by design of charge patterning in the intrinsically disordered RAM region of the Notch receptor. *Proc Natl Acad Sci U S A* 2017; 114: E9243–E9252.

(11) Harris J, Shadrina M, Oliver C, Vogel J, Mittermaier A. Concerted millisecond timescale dynamics in the intrinsically disordered carboxyl terminus of γ-tubulin induced by mutation of a conserved tyrosine residue. *Protein Sci* 2018; 27: 531–545.

(12) Duong VT, Chen Z, Thapa MT, Luo R. Computational Studies of Intrinsically Disordered Proteins. *J Phys Chem B* 2018; 122: 10455–10469.

(13) Meng F, Bellaiche MMJ, Kim JY, Zerze GH, Best RB, Chung HS. Highly Disordered Amyloid-β Monomer Probed by Single-Molecule FRET and MD Simulation. *Biophys J* 2018; 114: 870–884.

(14) Ouyang Y, Zhao L, Zhang Z. Characterization of the structural ensembles of p53 TAD2 by molecular dynamics simulations with different force fields. *Phys Chem Chem Phys* 2018; 20: 8676–8684.

(15) Fealey ME, Binder BP, Uversky VN, Hinderliter A, Thomas DD. Structural Impact of Phosphorylation and Dielectric Constant Variation on Synaptotagmin's IDR. *Biophys J* 2018; 114: 550–561.

(16) Grazioli G, Martin RW, Butts CT. Comparative exploratory analysis of intrinsically disordered protein dynamics using machine learning and network analytic methods. *Front Mol Biosci* 2019; 6: 42.

(17) Shrestha UR, Juneja P, Zhang Q, Gurumoorthy V, Borreguero JM, Urban V *et al.* Generation of the configurational ensemble of an intrinsically disordered protein from unbiased molecular dynamics simulation. *Proc Natl Acad Sci U S A* 2019; 116: 20446–20452.

(18) Shabane PS, Izadi S, Onufriev A V. General Purpose Water Model Can Improve Atomistic Simulations of Intrinsically Disordered Proteins. *J Chem Theory Comput* 2019; 15: 2620–2634.

(19) Robustelli P, Piana S, Shaw DE. The mechanism of coupled folding-upon-binding of an intrinsically disordered protein. *bioRxiv* 2020. doi:10.1101/2020.03.23.004283.

(20) Sridhar A, Orozco M, Collepardo-Guevara R. Protein disorder-to-order transition enhances the nucleosome-binding affinity of H1. *Nucleic Acids Res* 2020; 8: 5318–5331.

(21) Wang K, Ning S, Guo Y, Duan M, Yang M. The regulation mechanism of phosphorylation and mutations in intrinsically disordered protein 4E-BP2. *Phys Chem Chem Phys* 2020; 22: 2938–2948.

(22) Chen J, Im W, Brooks CL. Balancing solvation and intramolecular interactions: Toward a consistent generalized born force field. *J Am Chem Soc* 2006; 128: 3728–3736.

(23) Best RB, Hummer G. Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *J Phys Chem B* 2009; 113: 9004–9015.

(24) Best RB, Mittal J, Feig M, MacKerell AD. Inclusion of many-body effects in the additive CHARMM protein CMAP potential results in enhanced cooperativity of α-helix and β-hairpin formation. *Biophys J* 2012; 103: 1045–1051.

(25) Bottaro S, Lindorff-Larsen K, Best RB. Variational optimization of an all-atom implicit solvent force field to match explicit solvent simulation data. *J Chem Theory Comput* 2013; 9: 5641–5652.

(26) Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, De Groot BL *et al.* CHARMM36m: An improved force field for folded

and intrinsically disordered proteins. *Nat Methods* 2016; 14: 71–73.
(27) Lee KH, Chen J. Optimization of the GBMV2 implicit solvent force field for accurate simulation of protein conformational equilibria. *J Comput Chem* 2017; 38: 1332–1341.
(28) Robustelli P, Piana S, Shaw DE. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc Natl Acad Sci U S A* 2018; 115: E4758–E4766.
(29) Liu H, Song D, Zhang Y, Yang S, Luo R, Chen HF. Extensive tests and evaluation of the CHARMM36IDPSFF force field for intrinsically disordered proteins and folded proteins. *Phys Chem Chem Phys* 2019; 21: 21918–21931.

DR. JEAN MICHEL GRÉVY MACQUART
COORDINADOR DEL POSGRADO EN CIENCIAS
PRESENTE

Atendiendo a la solicitud para emitir DICTAMEN sobre la revisión de la TESIS titulada **"Estructura residual en la región intrínsecamente desordenada de Escargot",** que presenta la alumna **Teresa Hernández Segura (7220130508)** para obtener el título de **Doctor en Ciencias.**

Nos permitimos informarle que nuestro voto es:

NOMBRE	DICTAMEN	FIRMA
Dr. Carlos Daniel Amero Tello CIDC-UAEM	APROBADO	
Dr. Rodrigo Said Razo Hernández CIDC-UAEM	APROBADO	
Dr. Enrique Rudiño Piñera IBT-UNAM	APROBADO	
Dra. Verónica Mercedes Narváez Padilla CIDC-UAEM	APROBADO	
Dra. Laura Domínguez Dueñas FQ-UNAM	APROBADO	
Dr. Cesar Millán Pacheco FF-UAEM	APROBADO	
Dra. Carmen Nina Pastor Colón CIDC-UAEM	APROBADO	

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

VERONICA MERCEDES NARVAEZ PADILLA | Fecha:2020-11-24 18:55:50 | Firmante

HC3g4He8Bu8QXO2BjbMwFRyabGaBXgC/A/ihkYhsZd7Opa/UIZwyIAO95FZFM2tLg3OFwm8g/IhmaoRqyiVDij7MqJE8BfA6InOkb2EbfNYMnn5HEvU8sWjf9tiUMaTv9oCRbNEEL1tuC0E5OJ9XkdUxacAKUfeVS+UvHdpNrJ/xs6mKfzjJ8c+LdznbNuzrCBYGKbaeWXcbuTLO064Wcv0V3PKbgpl7LMBAe5OAcLWjOt+qWcdXBINox0ucnXm8jKCoEr6wDE651eUTPwE1YbyurP5K84v05Ulg2lyIGO5utaWSr2qUqca9LutlNB4GOnhfevngMrUsEEjKAGxLTQ==

RODRIGO SAID RAZO HERNANDEZ | Fecha:2020-11-24 19:06:48 | Firmante

IC3tzMpuVWV9ORqIn2oXNl3PUDNlXwmDOJn3ZyjWLzLg8vReQnDSniZKUjJmOWztBHIRZ56YWwWnfoICsPKS9ua3jIRFw/kcl1uUQz6W0WEoSGWryFDPiWYvSasA/kT83p3iNuRgmZmsuxm9hfaN9BesT4XHaFGwbBEmR8FV1IIN6XPdpGLKH8YlCJ6BqxAQUOD8GLldAEkcH7HpdMojKK0erMCCIMBQmuZPVgh1+lJU/4aiXVw/T2elJPQqJGDjHL7/WPHKnjrZii8XQziJT0qCgNVt+gXRxU1qjsHa1Rbtz7fYDyEnWOzw8CrTRGNvCMW0JF2Df4HGpI++BIINQw==

LAURA DOMÍNGUEZ DUEÑAS | Fecha:2020-11-24 19:20:49 | Firmante

vBnDOIxhTaH3qrl0oE/Xyy5ZotzSHT7bOY1V8+1IUNeyPKc26AQqrOiFWySc2sL7kunCDDbX90SjV+1jNtGa0rn+JIProHiC5vTw/2RQzXdhX5gTdgptd1sz4UTwGZ1Jx8xNQ8ML9dZz6RTFulb1+hJ05+mjLBswzVGtEHR1kgDSKdwZOEmVRtlhHjhPJmhcEnoZ5atCFjlJDSeVxlNIkdytHf7GOI0zUDYDv3J7m77uJ133p+96xURE8DlPrTDS/vqU8ApDbhuxrLf5FcTlOL/s3yLfyoAEEZjgFOGCG8SwQ/Licn0uq3XpXdNb4ep89ohCcZJX30gX2JvexJJ9oA==

CESAR MILLAN PACHECO | Fecha:2020-11-24 19:25:35 | Firmante

gfJyVXc4BODrgrsNrYs+qTQ9jdoPvQTODxiv3JjQ/ycMjZ/nG7z+OrcY7oNkc0Fg5Qvz+sXXaVMRFAIXw5/uxexnehNh7pwkhXlgMbY3DglxWqmXtG7HNZsDuq0pnqFf5YgbtxGBAb5uHptSquBJObulV3FpiacMT4Lkpa8TK4sfhXT74edFT63SMj9fUy9CdZhK9H7Uf+fTQtkYUgb3T3+9Jlz4/nVM5sXeAIKb/VEgODKRqQEj39kDGyKS29aflLMPjGBP02s4EEbd4jeFvhdp8ENgPNqupUIdftZOvPg38+MTTDZ2nnKbZg7vUagwn9TFvGYqK/xlzpO5vIEJlg==

CARLOS DANIEL AMERO TELLO | Fecha:2020-11-24 20:15:16 | Firmante

Ee19sObwgBrNPOQKvMyzTDuprsTJa1viz08S1vs1RVKAaBQ5hrY1tv2B5FVTW7ozuEx6r2JdTOcM+q2wcuNe5yEiuMY7JorKpp01qXLiuI5x3F7C8EqIHjxBoEn69lVFN7VGXUavVWpcuG0vJBV42lg0/Adcjstgtemh2gRqC06Notkio/SbKthMX4mRN5hAQPzHtA+x122SWVjJjcRMfNUaVDXFABGrG6fcYsoecYP0yqGTE9emP8CONX3BZbzkjeKezJkS+9n6aifQLwbKYuEpqK0zpDuI68r7mdlQGI7Q3E2HMjh+En0NqeaOhNWPdbgWeBJeL4TUhx+gM0nI0A==

ENRIQUE RUDIÑO PIÑERA | Fecha:2020-11-25 09:58:45 | Firmante

IMlWU9+9zjn1dH3WIISDiHTROnihnuHflQ2uL8U5ay2n+9xOyfONg4jKxwMU8hol4Al3GB0jOgTHTOErXmdLQybOrlSwD7GSq6IHwS1gnMUTi467yzvO3APjm3yDTf/1/CVszBUxxuG9foEVuE2pXhvhF2M60oLGhZWmXgHiyoA5d0zgDWKxE1b/Stc7UcX9K26/8C1WJLOh3GCHyor0wOKOaNutisbQfRqctpgGO4/bdCC1ntX8VQWdGr3MdgY9rFtFp6Dt2VLLehNFgc9mYV0tD+RDRTm5yO2tTLgLeWH0z36WfPi5p1i4VjRPMPCec8m5cGbapFb3sTzL1yyX1w==

CARMEN NINA PASTOR COLON | Fecha:2020-11-25 10:04:18 | Firmante

AO+wnTgLfE5sHg12KWQtFOmpsnxZMezNy2UFPLnzClPN4ozEirXEHyI/4KkJdMAh8BC/uOzbjp4rnCujWzlungvD+ExSgQbH/7yUQdA2vapXCFkpjiFKJBasDgn/QlPy8edYU+zItoljZfpzyqDYVyYlRYBhcgWyEg6aTwz9pLFQb4GxogDqk44qpaSCBTSA+0pJ2IHHGOJ+LJJebX54OIX1RTmbV98oxs8SSY8Z8ttV6NSyC/f5FGb4DjSpYR7HyfQkMHn9MBw5o084ygulWsvc3bMr6zyfW11/KAjZxw9/RqZZEJx3FJstd2XLQ2QbKfClR+y4j+BDbQwM+IjKNw==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

15H7lx

https://efirma.uaem.mx/noRepudio/FpyvaTITZD3U5pp6XtS2OJ1MFz4y2znm