Universidad Nacional Autónoma de México

Programa de Maestría y Doctorado en Ciencias Médicas Odontológicas y de la Salud.

División de Estudios de Posgrado e Investigación

Facultad de Odontología

Identificación de dominios de la unión a hidroxiapatita en la proteína del cemento 1 (CEMP1) mediante análisis bioinformáticos y simulaciones de dinámica molecular.

TESIS

Que para optar por el grado de:

Maestra en ciencias

Presenta:

CD. Diana Sofía Nolasquez Cruz

Tutor:

Dr. Eduardo Villarreal Ramírez

Universidad Nacional Autónoma de México

México, CDMX, Junio 2019

Dedicatorias

A mis padres:

Carmen Josefina Cruz López y Arturo Nolásquez Sánchez

Por ser los principales promotores de mis sueños, por su paciencia infinita y su amor incondicional.

A mis compañeros de clase y de laboratorio:
Lucía, Osmar, Febe, Laurita, José Luis, Paola, Maggie y Adrián.
Por qué dicen que los amigos son la familia que uno elige... gracias por su amistad.

A Íñigo Gaitán Salvatella
Por qué has estado conmigo en los momentos más difíciles, motivándome para seguir adelante. Gracias por creer en mí y brindarme tu apoyo de manera incondicional.

Si no hiciésemos cosas estúpidas nunca se haría nada inteligente.
Ludwig Wittgenstein.

Agradecimientos

Al Dr. Eduardo Villarreal Ramírez
Por aceptarme como su alumna y por su apoyo para la realización de este trabajo.

Al Dr. Marco Antonio Álvarez Pérez
Por abrirme las puertas del laboratorio de bioingeniería de tejidos del DEPeI de la Facultad de Odontología UNAM.

Al Dr. Higinio Arzate
Por permitirme realizar el estudio de dinámica molecular de CEMP1.

Al Dr. Luis Fernando lozano Aguirre Beltrán
Por la escritura de los programas en PERL.

A mi comité tutoral:
Dr. Jesús Ángel Arenas Alatorre
Dr. Higinio Arzate
Por su consejo y aportaciones durante mi periodo de maestría.

Institucionales

A la Universidad Nacional Autónoma de México
Máxima casa de estudios y mi segunda casa.

Se agradece el apoyo para la realización de este trabajo al proyecto PAPIIT/UNAM: IA207218 y a los recursos de cómputo de la supercomputadora Miztli proporcionados a través del proyecto: LANCAD-UNAM-DGTIC-324.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por la beca otorgada para mis estudios de maestría con el número de cvu 854964.

Índice

Resumen

El periodonto es una unidad anatómica de tejidos para unir el diente a la mandíbula y está compuesto por cuatro tejidos: la encía, el ligamento periodontal, el hueso alveolar y el cemento radicular. El cemento radicular es un tejido mineralizado especializado que cubre la superficie de las raíces del diente, carece de un aporte sanguíneo directo, inervación, drenaje linfático y no se somete a procesos de remodelación fisiológica. Las funciones del cemento radicular son proporcionar la unión del diente al hueso alveolar y mediar la inserción de las fibras del ligamento periodontal. Hay proteínas aisladas del cemento que se consideran proteínas cemento especificas, como la proteína del cemento 1 (CEMP1). La proteína CEMP1 tiene una secuencia de 247 aminoácidos, un peso molecular teórico de 26 kDa, con un punto isoeléctrico de 9.73 y una estructura secundaria compuesta por 61.5% de estructura aleatoria. CEMP1 se considera una proteína intrínsecamente desordenada (IDPs).

Se ha informado que CEMP1 es un regulador de la composición, deposición y morfología de los cristales de hidroxiapatita (HA) *in vitro*. Los estudios *in vivo* con hrCEMP1 purificado inducen la regeneración ósea en defectos de tamaño crítico en la calvaria de rata. Estos datos sugieren una posible acción terapéutica de esta proteína para la regeneración de tejidos mineralizados. Sin embargo, los mecanismos de acción de CEMP1 son aún desconocidos.

En este trabajo, presentamos un análisis bioinformático y simulaciones de dinámica molecular de las interacciones CEMP1 / HA. De acuerdo con nuestros resultados, las interacciones entre los péptidos derivados de CEMP1 se deben principalmente a fuerzas electrostáticas. Es importante destacar que nuestros resultados muestran que la fosforilación de proteínas desempeña un papel crucial en la alteración de la carga neta de CEMP1 y tuvo grandes efectos sobre la capacidad de los péptidos para interactuar con los cristales de hidroxiapatita.

Abstract

The periodontium is an anatomic unit of tissues to attach the tooth to the jaw and is composed of four tissues: gingiva, periodontal ligament, alveolar bone, and cementum. The cementum is a specialized mineralized tissue that covers the surface of the roots of the tooth, lacks direct blood supply, innervation, lymphatic drainage, and does not undergo physiological remodeling processes. The functions of the cementum are to provide the union of the tooth to the alveolar bone and to mediate the insertion of fibers of the periodontal ligament. There are proteins isolated from cementum that are considered cement-specific proteins, such as cementum 1 protein (CEMP1). The CEMP1 protein has a sequence of 247 amino acids, a theoretical molecular weight of 26kDa, with an isoelectric point of 9.73, and a secondary structure composed of 61.5% random structure. CEMP1 is considered an intrinsically disordered protein (IDPs).

CEMP1 has been reported as a regulator of the composition, deposition, and morphology of hydroxyapatite crystals (HA) in vitro. In vivo studies with purified recombinant human CEMP1 induces bone regeneration in critical size defects in rat calvaria. These data suggest a possible therapeutic action of this protein for the regeneration of mineralized tissues. Nevertheless, the mechanisms of CEMP1 actions are still unknown.

In this work, we present a novel bioinformatics analysis and molecular dynamics simulations of CEMP1/HA interactions. According to our results, the interactions between CEMP1-derived peptides are mainly due to electrostatic forces. Importantly, our results show that protein phosphorylation plays a crucial role in alteration the net charge of CEMP1 and had large effects on the ability of the peptides to interact with hydroxyapatite crystals.

Abreviaturas

2D Bidimensional

3D Tridimensional

ADP Adenosín difosfato

ADN Ácido desoxirribonucleico

ALP Fosfatasa alcalina (por sus siglas en inglés)

ANN Redes neuronales artificiales (por sus siglas en inglés)

ARN Ácido ribonucleico

ATP Adenosín trifosfato

BLAST Herramienta de búsqueda básica de alineación local.

BSP Sialoproteína ósea (por sus siglas en inglés)

DC Dicroísmo circular

CEMP1 Proteína del cemento 1

CFSSP Servidor de predicción de estructura secundaria Chou & Fasman

EGF Factor de crecimiento epidérmico (por sus siglas en inglés)

FGF Factor de crecimiento fibroblástico

GROMACS Máquina de Groningen para simulación química

HA Hidroxiapatita

hrCEMP1 Proteína recombinante humana de CEMP1

IDPs Proteínas intrínsecamente desordenadas (por sus siglas en inglés)

IDRs Regiones intrínsecamente desordenadas (por sus siglas en inglés)

kDa Kilo Dalton

SDM Simulaciones de Dinámica molecular

SS-NMR Resonancia magnética nuclear en estado sólido (por sus siglas en inglés)

NCBI Centro nacional para la información biotecnológica

ns Nano segundo

PDB Base de datos de proteínas (por sus siglas en inglés)

ps Pico segundo

Rg Radio de giro

NMR Resonancia magnética nuclear (por sus siglas en inglés).

RMSF Fluctuación de desviaciones cuadráticas medias

SOPMA Método de predicción auto-optimizado con alineación

Introducción

Periodonto

El periodonto (del latín peri que significa "alrededor" y griego odonto que significa diente) se define como el conjunto de tejidos de soporte y revestimiento de los órganos dentarios. Se compone de cuatro tejidos principales: encía, ligamento periodontal, cemento y hueso alveolar (Figura 1). Cada uno de los tejidos periodontales es distinto en ubicación, arquitectura, composición bioquímica y estos componentes juntos funcionan como una sola unidad funcional. El periodonto experimenta cambios con la edad y se encuentra sometido a cambios morfológicos relacionados con el ambiente bucal (Lindhe Jan et al., 2009; Newman Michael G. et al., 2018).

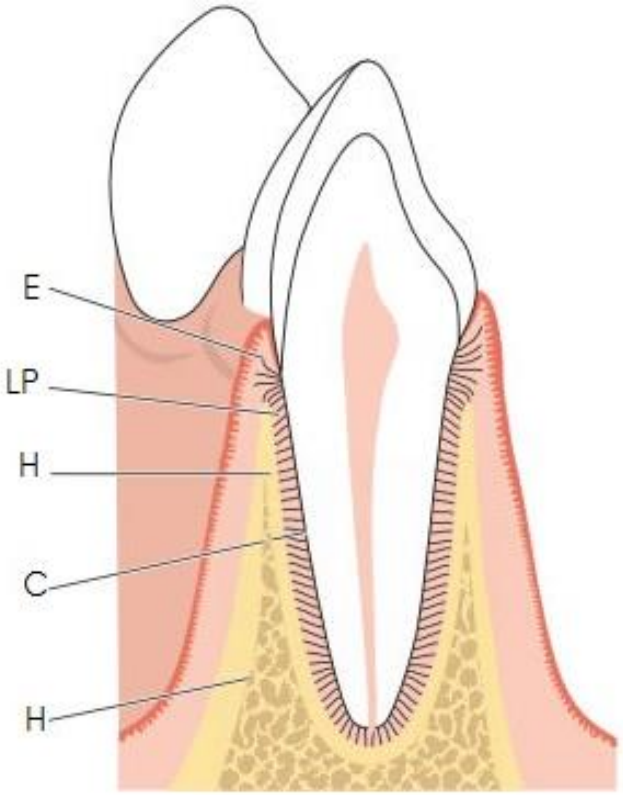


Figura 1. Tejidos periodontales: (E) encia, (LP) ligamento periodontal, (H) hueso alveolar, (C) cemento radicular. Tomado de http://oralcorp.com.ec/site/images/enfermedad.jpg Mayo 2017.

1. Encía

La mucosa bucal se continúa con la piel de los labios y con las mucosas del paladar blando y de la faringe. La mucosa bucal consta de:

- Mucosa masticatoria (incluye encía y la mucosa del paladar duro).
- Mucosa especializada (recubre la cara dorsal de la lengua).
- Mucosa de revestimiento (recubre las regiones restantes).

La encía está compuesta de una capa epitelio escamoso estratificado y tejido conjuntivo subyacente denominado lámina propia. Es la parte de la mucosa masticatoria que recubre la apófisis alveolar y rodea la porción cervical de los dientes hasta la unión cemento-esmalte. Anatómicamente está dividida en encía marginal, insertada o adherida e interdental, cada tipo de encía presenta diferencias considerables en diferenciación, histología y espesor de acuerdo a sus demandas funcionales (Newman Michael G. et al., 2018)

El componente tisular predominante en la encía es el tejido conjuntivo cuyos componentes principales son: fibras colágenas, reticulares y elásticas (60%) fibroblastos (5%) vasos, nervios y matriz (aproximadamente 35%). El tejido conectivo tiene componentes celulares y extracelular compuestos de fibras y sustancia fundamental. La sustancia fundamental llena el espacio entre las fibras y las células, es amorfa, está compuesto por

proteoglicanos y glicoproteínas y tiene un alto contenido de agua (Lindhe Jan et al., 2009; Newman Michael G. et al., 2018).

2. Ligamento periodontal

El ligamento periodontal está compuesto por un tejido conectivo vascular y altamente celular complejo que rodea la raíz del diente y se conecta a la pared interna del hueso alveolar. Es continuo con el tejido conectivo de la encía y se comunica con los espacios medulares a través de los canales vasculares en el hueso (Newman Michael G. et al., 2018)

El elemento más importantes del ligamento periodontal son las fibras principales, son haces de fibras colágenas dispuestas en seis grupos (fibras transeptales, alveolares, horizontales, oblicuas, apicales e inter-radiculares) y se desarrollan secuencialmente con el desarrollo de la raíz, las porciones terminales de las fibras principales se insertan en el cemento y el hueso se denominarán como fibras de Sharpey. Una vez incrustadas, las fibras de Sharpey se calcifican en un grado significativo debido a abundantes proteínas no colágenas que se encuentran típicamente en los huesos y en el cemento radicular. Estas proteínas contribuyen a la regulación de la mineralización (Lindhe Jan et al., 2009; Newman Michael G. et al., 2018).

Las funciones del ligamento periodontal se dividen en: físicas, formativas y de remodelación, nutricionales y sensoriales (Newman Michael G. et al., 2018).

Funciones físicas:

- Provisión de una "cubierta" de tejido blando para proteger los vasos y los nervios de lesiones por fuerzas mecánicas.
- Transmisión de las fuerzas de oclusión al hueso.
- Fijación de los dientes al hueso (la movilidad dental está determinada en gran medida por el espesor, la altura y la calidad del ligamento periodontal).
- Mantenimiento de los tejidos gingivales en su correcta relación con los dientes.

Funciones formativas y de remodelación:

Las células del ligamento periodontal participan en la formación de cemento y hueso alveolar, el cual ocurre durante los movimiento fisiológicos del diente, durante el acoplamiento del periodonto a las fuerzas oclusales y durante la reparación de lesiones.

Funciones nutricionales y sensoriales:

El ligamento periodontal suministra nutrientes al cemento, hueso alveolar y la encía a través de los vasos sanguíneos; además presenta abundantes fibras nerviosas con baroreceptores y nociceptores capaces de la transmisión de sensaciones táctiles, de presión y de dolor a través de las vías trigeminales (Newman Michael G. et al., 2018).

3. Hueso alveolar

El proceso alveolar es la parte de los maxilares superior e inferior que forma y sostiene los alveolos de los dientes, se forma cuando el diente entra en erupción para proporcionar el enlace óseo al ligamento periodontal en formación (fibras de Sharpey). Junto con el cemento y el ligamento periodontal constituyen el aparato de inserción del diente, cuya función principal consiste en distribuir y absorber las fuerzas generadas por la masticación. (Lindhe Jan et al., 2009).

Los procesos alveolares se desarrollan y someten a remodelación junto con la formación y erupción de los dientes, de modo que después de la extracción de los mismos se reabsorbe lentamente hacia el cuerpo de hueso maxilar y/o mandibular. Por lo tanto, el tamaño, la forma, la ubicación y la función de los dientes determinan su morfología (Newman Michael G. et al., 2018).

El hueso está formado por dos tercios de materia inorgánica y un tercio de matriz orgánica. La materia inorgánica está compuesta principalmente por los minerales calcio y fosfato, junto con el hidroxilo, carbonato, citrato y trazas de otros iones como el sodio, el magnesio y el flúor. Los cristales de hidroxiapatita (HA) tienen la

siguiente fórmula $[Ca_{10}(PO_4)_6(OH)_2]$. La matriz orgánica consiste principalmente en colágena tipo I (90%) y un 10% de proteínas no colágenas como la osteocalcina, la osteonectina, la proteína morfogenética ósea, fosfoproteínas y proteoglicanos (Newman Michael G. et al., 2018).

4. Cemento radicular

El cemento radicular fue descrito por primera vez por Purkinje et al. en 1835 (Foster, 2017) y hasta hace poco ha permanecido como un tejido mal definido a nivel celular y molecular (Saygin et al., 2000). Es un tejido mineralizado único dentro del cuerpo que recubre las superficies radiculares de los órganos dentarios; a diferencia del tejido óseo, no contiene vasos sanguíneos ni linfáticos, carece de inervación y no experimenta procesos de remodelación o resorción fisiológica (Lindhe Jan et al., 2009; Saygin et al., 2000). A pesar de estas diferencias, el cemento es muy similar al hueso (Saygin et al., 2000).

Dentro de las funciones del cemento radicular están: proporcionar la unión del diente al hueso alveolar mediante la inserción de fibras del ligamento periodontal; prevenir la reabsorción de la raíz durante la remodelación del periodonto (Hughes, 2015); actúa como una barrera para delimitar el crecimiento epitelial que puede perjudicar la unión del diente y la presencia de una capa de cemento continua es necesaria para actuar como barrera microbiana (Arzate et al., 2015).

Las fibras de colágena presentes en la matriz del cemento presentan dos distintos orígenes, las fibras extrínsecas cuya procedencia es a partir de los fibroblastos de las fibras de Sharpey, y las fibras intrínsecas que son sintetizadas por los cementoblastos. Los cementoblastos también sintetizan proteínas no colágenas de la sustancia fundamental interfibrilar, como los proteoglicanos, glicoproteínas y fosfoproteínas (Newman Michael G. et al., 2018)

La deposición de cemento es un proceso continuo que se realiza a velocidades variables a lo largo de la vida. La formación del cemento es importante para la apropiada maduración del periodonto, tanto durante el desarrollo como durante la regeneración de tejidos periodontales (Newman Michael G. et al., 2018; Saygin et al., 2000). Se cree que una variedad de factores de crecimiento, como el factor de crecimiento derivado de plaquetas, los factores de crecimiento insulínicos, el factor de crecimiento transformante beta 1, el factor de crecimiento fibroblástico básico, la dexametasona y proteínas morfogéneticas óseas; se producen durante la formación del cemento y luego se almacenan en la matriz de cemento para inducir la regeneración del ligamento periodontal cuando sea necesario (Arzate et al., 2015)

El cemento se compone de una fase orgánica y una fase inorgánica o mineral. La fase inorgánica representa entre el 45-50% del peso y está compuesta por HA. La fase orgánica está compuesta por proteínas colágenas y no colágenas (Arzate et al., 2015; Lindhe Jan et al., 2009; Newman Michael G. et al., 2018; Saygin et al., 2000).

- El principal tipo de colágena es de tipo I (90%) y desempeña un papel estructural durante el proceso de biomineralización, sirviendo como un reservorio para la nucleación de HA; la colágena tipo III (5%) cubre las fibras de colágena tipo I.

- Proteínas no colágenas como: glicosaminoglicanos (ácido hialurónico, sulfato de dermatán, sulfato de condroitina y sulfato de queratán) que presentan una distribución diferencial en el cemento lo que sugiere que pueden desempeñar papeles distintos durante el proceso de cementogénesis además de regular la biomineralización del cemento; fosfoproteínas (osteopontina y sialoproteína ósea) que se han postulado como reguladores de la nucleación y crecimiento del cristal de HA; proteínas Gla (proteína matriz ácido gamma-carboxiglutámico y osteocalcina) que tienen alta afinidad por Ca^{2+} y HA a través de la interacción con el residuo de Gla y cuyo papel se ha asociado con la regulación de la mineralización; fosfatasa alcalina que está altamente expresada en las células del ligamento periodontal, funciona como un inhibidor de la formación de hidroxiapatita indicando que desempeña un papel biológico clave en la mineralización del hueso y el cemento (Arzate et al., 2015; Lindhe Jan et al., 2009; Newman Michael G. et al., 2018; Saygin et al., 2000).

Factores moleculares asociados al cemento.			
Actividad propuesta	**Desarrollo del cemento**	**Maduración del cemento**	**Regeneración del cemento**
Adhesión o quimioatracción	Proteoglicanos Osteopontina Sialoproteína ósea Fibronectina Laminina No establecidas Colágena I, III, XII No establecida Factores de la vaina epitelial de Hertwig	Proteoglicanos Osteopontina Sialoproteína ósea Fibronectina No establecidas Proteína de adhesión del cemento Colágena I, III, XII Tenascina Factores de la vaina epitelial de Hertwig	No establecidas Osteopontina Sialoproteína ósea No establecidas No establecidas No establecidas Colágena I, III No establecidas No establecidas
Mitogénesis	Hormona del crecimiento Factor de crecimiento transformante-β Factor de crecimiento tipo insulínico I	No establecidos Factor de crecimiento transformante-β Factor de crecimiento del cemento/ Factor de crecimiento tipo insulínico I	No establecidas No establecidas No establecidas
Diferenciación	Proteína relacionada con la hormona paratiroidea Factor de crecimiento transformante-β Proteínas morfogenéticas óseas Factores de la vaina epitelial de Hertwig Factor de transcripción específico de osteoblastos	No establecidos Factor de crecimiento transformante-β Proteínas morfogenéticas óseas Factores de la vaina epitelial de Hertwig Factor de transcripción específico de osteoblastos	No establecidas No establecidas No establecidas No establecidas No establecidas
Mineralización	Sialoproteína ósea Osteocalcina Osteopontina Colágena I, XII Proteoglicanos	Sialoproteína ósea Osteocalcina Osteopontina Colágena I, XII Proteoglicanos	Sialoproteína ósea No establecidas Osteopontina Colágena I No establecidas

Tabla 1 Factores moleculares asociados al desarrollo, maduración y regeneración del cemento. Modificado de Saygin et al., 2000.

En la tabla 1 se muestran los factores moleculares asociados al desarrollo, maduración y regeneración del cemento (algunos factores son sugeridos y aún están bajo investigación) (Saygin et al., 2000).

5. Proteínas Cemento Específicas.

La matriz extracelular de diferentes tejidos comparte muchas similitudes y, sin embargo, tiene diferentes propiedades funcionales que las hacen únicas. Estas propiedades pueden ser el resultado de diferencias cuantitativas y/o cualitativas entre sus componentes. Durante años se creyó que diferentes tejidos mineralizados contienen moléculas específicas que no están presentes en ningún otro tejido, sin embargo estas moléculas también se expresan en otros tejidos, aunque a concentraciones considerablemente menores, y por lo tanto podían considerarse todavía como marcadores específicos. Existen varias proteínas que se han aislado del cemento y se consideran proteínas cemento-específicas (Arzate et al., 2015).

5.1 Proteína del cemento 1 (CEMP1)

La proteína de cemento 1 (CEMP1) fue aislada a partir un cementoblastoma humano (Arzate et al., 2002). El gen que codifica para esta proteína se localiza en el brazo corto del cromosoma 16 en el locus 13.3 (16p13.3). El producto del gen de la proteína CEMP1 está enriquecido en prolina (11,3%), glicina (10,5%), alanina (10,1%), serina (8,9%), leucina (8,1%), treonina y arginina (cada uno 7,7%) y contiene bajos niveles de triptófano, ácido aspártico, isoleucina (cada uno 2,0%), fenilalanina (1,6%) y no cuenta con tirosinas. La secuencia de aminoácidos indica que la proteína CEMP1 probablemente sea una proteína nuclear; sin embargo, no tiene motivos de unión al ADN (Arzate et al., 2015).

Consta de una secuencia de 247 aminoácidos con un peso molecular calculado de 26 kDa y después de modificaciones posteriores a la traducción aumente a 50 kDa (Álvarez-Pérez et al., 2006). Se ha reportado como una proteína alcalina con un punto isoeléctrico de 9,73, sin péptido señal. El análisis de dicroísmo circular

muestra que su estructura secundaria está compuesta por un 28,6% de hélice, un 9,9% de lámina β y un 61,5% de estructura aleatoria (Romo-Arévalo et al., 2016), este alto porcentaje de estructura aleatoria se asocia con proteínas intrínsecamente desordenadas (IDPs) que han demostrado ser multifuncionales y tener diversos propiedades de unión.

La proteína CEMP1 se encuentra altamente expresada en cementoblastos y células progenitoras del ligamento periodontal (Álvarez-Pérez et al., 2006); experimentos *in vitro* han demostrado que promueve la diferenciación (Arzate et al., 1996; Komaki et al., 2012) y la transfección de CEMP1 en células no osteogénicas resulta en una diferenciación hacia un fenotipo mineralizante (Carmona-Rodríguez et al., 2007). Además se ha reportado como un regulador de la composición, deposición y morfología de los cristales de hidroxiapatita (HA) (Álvarez Pérez et al., 2003) y estudios *in vivo* han demostrado que CEMP1 recombinante humana (hrCEMP1) produce la regeneración ósea en defectos de tamaño critico en calvaria de ratas (Serrano et al., 2013)

Todos estos datos antes mencionados sugieren una posible acción terapéutica de esta proteína en el tratamiento de defectos óseos, así como en la regeneración de otros tejidos mineralizados como el cemento radicular.

Fundamentos de la metodología

Análisis bioinformático

La investigación que se realiza dentro del campo del área biológica provee de grandes cantidades de información que crece de manera exponencial en tamaño y complejidad, debido a la disponibilidad de estas enormes cantidades de datos biológicos y la necesidad de transformar estos datos en información biológica útil, es que se ha desarrollado el campo de la bioinformática, con la generación de técnicas computacionales avanzadas que permiten el procesamiento y análisis eficiente de los datos (Escobar et al., 2011).

La bioinformática es el resultado de la conjugación de varias disciplinas, entre las cuales están la informática, las matemáticas, la estadística, la química y las ciencias biológicas que junto con técnicas computacionales permiten analizar un gran volumen de datos biológicos sobre el ADN, ARN y proteínas; por lo tanto la finalidad de la bioinformática es el desarrollo de bases de datos para almacenar y recuperar datos biológicos, algoritmos para analizar y determinar sus relaciones y las herramientas estadísticas para identificar e interpretar los datos (Escobar et al., 2011).

Las herramientas que facilitan la investigación bioinformática se pueden clasificar en cuatro clases:

a) Algoritmos de recuperación de datos
   Proporcionan un acceso a una amplia gama de dominios de datos incluyendo secuencias de nucleótidos y proteínas reportadas en la literatura, genomas completos, estructuras 3D y más.
b) Comparación de la secuencia y herramientas de alineación.
c) Algoritmos para el descubrimiento de patrones
   Se utilizan para buscar patrones o características de los datos, básicamente lo que se pretende es encontrar grupos en un determinado conjunto de datos, de tal manera que los objetos en el mismo grupo sean similares entre sí y diferentes a los de otros grupos.
d) Herramientas de visualización.
   Permiten una visualización interactiva y gráfica de los datos genómicos

La generación de grandes cantidades de información, a partir del desarrollo de la tecnología de secuenciación de ADN en la década de 1970, ha ejercido un papel importante en la investigación biomédica y ha proporcionado vínculos entre diversos sistemas biológicos. Dichas conexiones han acelerado el progreso de la investigación, donde la búsqueda de homología en bases de datos de secuencias de aminoácidos y ADN, se ha convertido en un método común y en continuo crecimiento. (Altschul et al., 1994).

Actualmente se han secuenciado los genomas de varios organismos como por ejemplo: la levadura *Saccharomyces cerevisiae*, la bacteria *Escherichia coli*, el gusano *Caenorhabditis elegans* y la mosca de fruta *Drosophila melanogaster*, por mencionar algunas especies relevantes. Dichos estudios han generado una gran cantidad de información en secuencias genéticas e información sobre la función biológica de las mismas, que en general se han usado para predecir la función de genes similares en otros organismos. (David W. Mount, 2001).

Dentro del análisis bioinformático, se incluye la comparación de la secuencia de ADN para el descubrimiento de similitudes funcionales y/o estructurales entre múltiples secuencias biológicas, incluye la alineación de secuencias, la búsqueda en base de datos de secuencias, el descubrimiento de patrones, la reconstrucción de las relaciones evolutivas y la formación y comparación del genoma (Escobar et al., 2011).

1. Alineamiento de secuencias

El alineamiento de secuencias es una manera de representar y comparar dos o más secuencias de ADN, ARN o de estructura primaria de proteínas, mediante la búsqueda de patrones de caracteres comunes, con el propósito de establecer un grado de similitud de secuencias que corresponde a un porcentaje de residuos alineados, similares en propiedades físico-químicas como tamaño, carga eléctrica o hidrofobicidad (Escobar et al., 2011).

Cuando dos secuencias son descendientes de un origen evolutivo común se dice que tienen una relación homóloga u homología, si se puede establecer una relación de homología, es probable que las secuencias

hayan mantenido la misma función, estructura y/o actividad bioquímica, y que divergieron unas de otras durante el proceso evolutivo (Escobar et al., 2011).

Para poder analizar la gran cantidad de información generado de manera eficiente y certera, se han escrito distintos algoritmos para poder leer las secuencias en formato de texto y realizar búsquedas en las bases de datos, así como evaluar la calidad de los resultados arrojados. La efectividad de las búsquedas en la base de datos depende de un gran número de factores correlativos. (Altschul et al., 1994). Entre estos se incluyen los siguientes:

- Medida máxima del par de segmentos
  Las medidas de similitud de secuencia generalmente pueden clasificarse como globales o locales (Figura 2).
  - Los algoritmos de similitud global optimizan la alineación general de dos secuencias, que pueden incluir grandes extensiones de baja similitud.
  - Los algoritmos de similitud local solo buscan subsecuencias relativamente conservadas, y una sola comparación puede producir varias alineaciones de subsecuencias distintas.

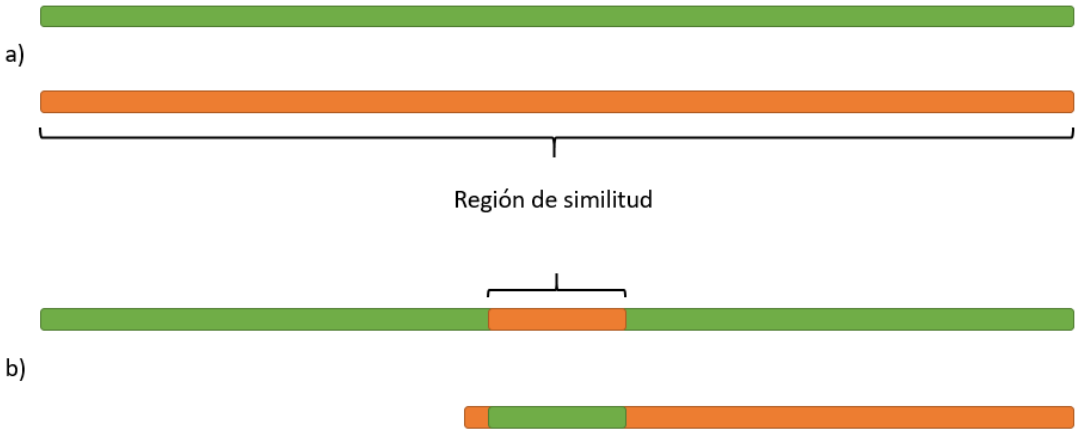


Figura 2. Medidas de similitud de secuencia. a)Global b) Local.

- Sistemas de puntuación
  La mayoría de los algoritmos de búsqueda de bases de datos clasifican las alineaciones por puntuación, cuyo cálculo depende de un sistema de puntuación particular. Generalmente hay un sistema predeterminado, pero que puede o no ser el ideal dependiendo del problema en particular. Por lo tanto, un programa de búsqueda de base de datos debería hacer disponibles una variedad de sistemas de puntuación.
- Bases de datos
  Una base de datos de secuencias bien descritas, permitirá que los resultados brinden coincidencias de datos con mayor precisión y con menores errores de secuenciación; el uso de una base de datos de secuencias adecuada y actualizada es un elemento vital para la búsqueda de similitud.
- Valor esperado
  El Valor esperado para las coincidencias encontradas indica la validez de la coincidencia, cuanto menor sea el valor esperado, más probable es que la coincidencia sea buena y presente una similitud real, en lugar de una coincidencia aleatoria. (McGinnis and Madden, 2004).
- Algoritmos y programas
  Los primeros estudios de comparación de secuencias se enfocaron en la alineación de secuencias completas (globales). Sin embargo, con el reconocimiento de que las proteínas a menudo solo comparten regiones aisladas de similitud, la atención se dirigió a algoritmos para la alineación local. Las alineaciones locales se evalúan mediante un puntaje, que se calcula como la suma de puntajes para pares alineados de residuos y puntajes para gaps. En general, las consideraciones relevantes al elegir un

algoritmo particular son los requisitos de hardware, la velocidad y la sensibilidad a las relaciones biológicas. La idea de optimizar una medida de similitud local es común a prácticamente todos los programas populares, y los resultados que producen, por lo tanto, no difieren de ninguna manera verdaderamente esencial.

- Estadísticas de alineación local
  No todas las relaciones de secuencia biológicamente importantes se detectarán mediante programas de búsqueda de similitud de secuencia e incluso cuando se encuentren, pueden perderse entre similitudes irrelevantes o aleatorias. Si bien la parte experimental es el árbitro final de importancia biológica, el análisis matemático puede indicar qué similitudes es poco probable que hayan surgido por casualidad.
- Matrices de puntaje y costo de gaps
  Se han desarrollado diferentes matrices de puntuación de sustitución de aminoácidos para su uso con programas de comparación de secuencias y búsqueda de bases de datos y se han utilizado una variedad de fundamentos para su construcción. Sin embargo, es posible mostrar que, en el contexto de la búsqueda de pares de segmentos de puntuación alta sin gaps, cualquier matriz de este tipo tiene una distribución de frecuencia de pares de aminoácidos implícita que caracteriza las alineaciones que está optimizada para encontrar. Supóngase que "P" es la frecuencia con la que aparece el aminoácido "i" en las secuencias de proteínas y, dentro de la clase de alineaciones buscadas, sea "q" la frecuencia con la que los aminoácidos "i" y "j" están sincronizados. Luego, los puntajes que mejor distinguen estas alineaciones del azar están dados por la fórmula:

$$Sij = \log \frac{qij}{Pi\ Pj} \tag{1}$$

Se han desarrollado diferentes métodos para estimar las frecuencias con las que los distintos aminoácidos tienden a mutarse; algunos de estos métodos de basan en la evolución molecular para calcular las frecuencias objetivo y matrices de puntuación correspondientes, como lo es el caso de las matrices "PAM". Un enfoque diferente para la estimación de las frecuencias objetivo apropiadas no se basa en el ajuste de un modelo evolutivo, sino en la observación directa de alineamientos de secuencias relativamente distantes, pero presumiblemente correctos, como es el caso de las matrices "BLOSUM".

La teoría que vincula las matrices de sustitución con las frecuencias objetivo se establece solo para las alineaciones locales que carecen de gaps. Por lo tanto, las alineaciones con gaps presentan el problema adicional de elegir los costos adecuados para los gaps. Los algoritmos más simples requieren que estos costos sean una función lineal de la longitud de los gaps, pero también hay algoritmos eficientes para costos de gaps más generales. Debido a que no existe una teoría, los costos de gaps apropiados generalmente se han elegido por ensayo y error.

1.1 BLAST

La herramienta de búsqueda básica de alineación local (BLAST), es un programa que se puede utilizar a través de una interfaz web o como una herramienta independiente (McGinnis and Madden, 2004). Es de dominio público y puede utilizarse gratuitamente desde el servidor del Centro Nacional para Información Biotecnológica (NCBI). El programa BLAST es utilizado para la predicción funcional de genes o secuencias proteicas. Por lo general, cuando una nueva secuencia es obtenida, se usa el programa para compararla con otras secuencias que han sido previamente caracterizadas, para así poder inferir su función.

BLAST realiza alineaciones "locales" puesto que la mayoría de las proteínas son de naturaleza modular, y los dominios funcionales a menudo se repiten dentro de la misma proteína, así como a través de diferentes proteínas de diferentes especies, BLAST hace la búsqueda de estos dominios o tramos más cortos de similitud de secuencia.

Cuando se envía una consulta, la secuencia más cualquier otra información de entrada, como la base de datos a buscar, el tamaño de la palabra, el valor esperado, etc., se envían al algoritmo del programa. BLAST funciona haciendo primero una tabla de consulta de todas las "palabras" (subsecuencias cortas) en la secuencia de consulta y la base de datos de secuencias se escanea en busca de estas "palabras". (AltschuP et al., 1990).

Una vez que el programa ha encontrado una secuencia similar a la consulta en la base de datos, es útil tener una idea de si la alineación es "buena" y si representa una posible relación biológica, o si la similitud observada es atribuible solo al azar; por lo que BLAST genera una puntuación y un valor esperado (valor E) para cada par de alineación. La puntuación es una medida de similitud y un indicador de cuan buena es una alineación; cuanto mayor sea la puntuación, mejor será la alineación. En términos generales, esta puntuación se calcula a partir de una fórmula que toma en cuenta la alineación de residuos similares o idénticos, así como cualquier gap introducida para alinear las secuencias. (Madden, 2002).

Un elemento clave en este cálculo es la "matriz de sustitución", que asigna una puntuación para alinear cualquier posible par de residuos. BLAST utiliza el algoritmo de Smith-Waterman (Smith and Waterman, 1981). El cual asigna puntajes a las inserciones, deleciones y sustituciones; las identidades y los reemplazos conservadores tienen puntajes positivos, mientras que los reemplazos poco probables tienen puntajes negativos.

Para las comparaciones de secuencias de aminoácidos las matrices BLOSUM62 y PAM-120 son la opción predeterminada para la mayoría de los programas BLAST; los programas que realizan comparaciones de nucleótidos vs nucleótidos, no utilizan matrices específicas de proteínas, por lo tanto, para las comparaciones de secuencias de ADN se puntúan identidades +5, y desajustes – 4. Se define el par con la puntuación más alta de grado de similitud de segmentos de la misma longitud llamado segmento par máximo (MSP); cuando se identifica una coincidencia, esta se utiliza para inicial extensiones sin y con gaps, y una subsecuncia se define como un MSP si su puntaje no puede mejorarse al extender o acortar ambos segmentos. (AltschuP et al., 1990).

Al buscar en una base de datos de miles de secuencias, generalmente solo algunas, si existen, serán homólogas a la secuencia de consulta. Por lo tanto, solo es de interés identificar las entradas de secuencia con puntuaciones de MSP sobre alguna puntuación de corte "S". Para acelerar las búsquedas en la base de datos, BLAST minimiza el tiempo dedicado a las regiones de secuencia cuya similitud con la consulta tiene pocas posibilidades de superar esta puntuación. Deja que un par de palabras sea un par de segmentos de longitud fija w. La estrategia principal de BLAST es buscar solo pares de segmentos que contengan un par de palabras con una puntuación de al menos T. Al escanear una secuencia, se puede determinar rápidamente si contiene una palabra de longitud w que pueda emparejarse con la secuencia de consulta para producir un par de palabras con un puntaje mayor o igual que el umbral T, cualquiera que cumpla con este requisito se extiende para determinar si está contenido dentro de un par de segmentos cuyo puntaje es mayor o igual a S. Cuanto menor sea el umbral T, mayor será la probabilidad que un par de segmentos con una puntuación de al menos S contendrá un par de palabras con una puntuación de al menos T. Un pequeño valor para T, sin embargo, aumenta el número de segmentos y, por lo tanto, el tiempo de ejecución del algoritmo. (AltschuP et al., 1990).

Básicamente la implementación del algoritmo corresponde a tres pasos que varían dependiendo de si la base de datos contiene proteínas o secuencias de ADN.

- Compilar una lista de palabras de puntuación alta.
- Escanear la base de datos en busca de resultados.
- Extender los resultados.

Existen diversos tipos de BLAST disponibles para las diferentes comparaciones de secuencias de acuerdo al tipo de secuencia de interés y la base de datos utilizada, por ejemplo, una consulta de ADN a una base de datos de ADN, una consulta de proteínas a una base de datos de proteínas y una consulta de ADN, traducida en los seis marcos de lectura, a una base de datos de secuencias de proteínas; las variantes de BLAST se resumen en la tabla 2.

Páginas de búsqueda	Combinación de consulta y base de datos	Tipo de alineación	Programas y funciones.
Blast de nucleótidos	Nucleótido vs Nucleótido	Nucleótido vs Nucleótido	Megablast: para identificación de secuencia, comparación entre especies. Megablast discontinuo: para comparar especies cruzadas, buscar con secuencias de codificación Blastn: para buscar con consultas más cortas, comparación entre especies.
Blast Proteico	Proteína vs Proteína	Proteína vs Proteína	Blastp: Identificación de secuencia general y búsquedas de similitud. DELTA-BLAST: búsqueda de similitud de proteínas con mayor sensibilidad que blastp. PSI-BLAST: búsqueda iterativa para la construcción de matriz de puntuación específica de posición (PSSM) o identificación de parientes distantes para una familia de proteínas PHI-BLAST: alineación de proteínas con patrón de entrada como ancla / restricción
Blastx	Nucleótido (traducido) vs Proteína	Proteína vs Proteína	blastx: para identificar productos proteicos potenciales codificados por una consulta de nucleótidos
Tblastn	Proteína vs Nucleótido (traducido)	Proteína vs Proteína	tblastn: para identificar secuencias de bases de datos que codifican proteínas similares a la consulta
Tblastx	Nucleótido (traducido) vs Nucleótido (traducido)	Proteína vs Proteína	tblastx: para identificar secuencias de nucleótidos similares a la consulta en función de su potencial de codificación

Tabla 2. Características principales de las páginas de búsqueda de BLAST en la categoría "BLAST básico". En fondo gris la página de búsqueda utilizada y en letra roja el programa utilizado.

La página de inicio de BLAST enumera las variedades de búsquedas de por tipo: Nucleótido, Proteína, Traducción y Genomas en la figura 3 se muestran los parámetros por defecto para cada uno.

	Expect value	Word size	Reward match	Penalty mismatch	Gap existence	Gap extension	Percentage identity	Filtering	Matrix
Nucleotide									
Discontiguous megaBLAST	10	11	1	−2	0	2.5	None	Low complexity; mask for lookup table	–
MegaBLAST	10	28	1	−2	0	2.5	None	Low complexity	–
Standard BLASTN	10	11	1	−3	5	2	–	Low complexity	–
Short Nucleotide Sequences	1000	7	1	−3	5	2	–	None	–
Protein									
Standard blastp	10	3	–	–	11	1	–	Low complexity	BLOSUM62
Psi-BLAST	10	3	–	–	11	1	–	None	BLOSUM62
Phi-BLAST	10	3	–	–	11	1	–	None	BLOSUM62
Short Protein Sequences	20 000	2	–	–	11	1	–	None	PAM30
RPS-BLAST	10	3	–	–	11	1	–	Low complexity	BLOSUM62
Translated									
Blastx	10	3	–	–	11	1	–	Low complexity	BLOSUM62
Tblastn	10	3	–	–	11	1	–	Low complexity	BLOSUM62
Tblastx	10	3	–	–	11	1	–	Low complexity	BLOSUM62
Special Pages									
Genome BLAST pages (blastn)	0.01	28	1	−2	0	0	None	Low complexity; Human repeat	–

Figura 3. Lista de los diferentes enlaces disponibles en la página de NCBI BLAST y los párametros por defecto de cada enlace, tomado de McGinnis and Madden, 2004.

Para la búsqueda de nucleótido vs nucleótido, la velocidad y sensibilidad varían con el tamaño de la palabra y el tipo de extensión de los gaps. El programa más rápido es megaBLAST, que por defecto tiene un tamaño de palabra grande (se requiere de una coincidencia exacta de 28 bases para iniciar la extensión) y un algoritmo de extensión que no genera un costo a la existencia de gaps, sino un costo de extensión de gaps, por lo que es ideal para comparar secuencias similares, por ejemplo, del mismo organismo.

La manera en que se utiliza BLAST es ingresar una secuencia de nucleótidos o proteínas como una consulta en todas (o un subconjunto de) las bases de datos públicas de secuencias, pegando la secuencia en el cuadro de texto en formato FASTA o con identificador de secuencia (por ejemplo GenBank). Esto envía la consulta a través de Internet, la búsqueda se realiza en las bases de datos y servidores del NCBI, después de que el algoritmo ha buscado todas las "palabras" posibles de la secuencia de consulta y las ha extendido al máximo, reúne la mejor alineación para cada par de consulta-secuencia y escribe esta información en una estructura de datos SeqAlign, esta estructura en sí misma no contiene la información de secuencia; más bien, se refiere a las secuencias en la base de datos. El BLAST Formatter, que se encuentra en el servidor BLAST, puede usar la información en SeqAlign para recuperar las secuencias similares encontradas y mostrarlas de varias maneras. Por lo tanto, una vez que se ha completado una consulta, los resultados se pueden reformatear sin tener que volver a ejecutar la búsqueda y se generan en el navegador en el formato de visualización elegido, el programa permite ver los resultados en varios formatos, que incluyen la tabla de resultados simplificada, XML y el reporte "clásico" de BLAST que consta de tres secciones principales: 1) el encabezado, que contiene información sobre la secuencia de consulta, la base de datos buscada y una descripción gráfica; 2)) las descripciones de una línea de cada secuencia de base de datos que coinciden con la secuencia de consulta; 3) las alineaciones para cada secuencia de base de datos coincidente, ya que puede haber más de una alineación para una secuencia de base de datos que coincida. (Madden, 2002).

2. Análisis de secuencias de aminoácidos.

2.1 Estructura de proteínas.

Las proteínas constituyen los elementos estructurales de las células y además ejecutan prácticamente todas las funciones celulares (Alberts Bruce et al., 2016). Químicamente están compuestas por una o varias cadenas no ramificadas de aminoácidos unidos entre sí mediante enlaces peptídicos, por ello también reciben el nombre de polipéptidos.

Los aminoácidos están compuestos de un carbono quiral (presenta sus cuatro sustituyentes diferentes, con excepción de la glicina) unido a un grupo amino, un grupo carboxilo, un hidrógeno y una cadena lateral o radical (R) de estructura variable que determina la identidad y propiedades químicas de cada uno de los aminoácidos. Existen veinte diferentes aminoácidos codificados en el ADN, que pueden unirse en diferente orden y formar cadenas de diferente longitud, de tal manera que cada proteína tiene una secuencia única de aminoácidos que se conoce como estructura primaria.

La secuencia repetida de átomos a lo largo de la cadena se conoce como esqueleto polipeptídico o "backbone". Unidas a esta cadena repetitiva están aquellas partes de los aminoácidos que no forman parte del enlace peptídico (cadenas laterales). Las diferentes propiedades de los aminoácidos que componen a una proteína hace que al encontrarse en un medio acuoso se plieguen; es decir algunos aminoácidos con carga buscan interactuar con el agua, mientras que aminoácidos hidrofóbicos (sin carga) evitan este contacto; de esta manera los aminoácidos quedan más cercanos unos con otros e interaccionan entre sí. Debido a el radio de van der Waals de los átomos se limita el número posible de ángulos de enlace en una cadena polipeptídica (impedimentos estéricos), reduciendo el número de disposiciones tridimensionales (o conformaciones) de los átomos. A pesar de ello, una proteína puede plegarse en un número enorme de conformaciones diferentes (Alberts Bruce et al., 2016).

El plegamiento de una proteína está determinado por muchos tipos de enlaces no covalentes (enlaces de hidrógeno, las interacciones electrostáticas y las fuerzas de van der Waals), formados entre partes de la misma cadena polipeptídica (implica átomos del "backbone", así como átomos de las cadenas laterales); si bien los

enlaces no covalentes son mucho más débiles que los enlaces covalentes, actuando en paralelo pueden mantener unidas dos regiones y la fuerza combinada de una gran número de estos enlaces estabiliza el siguiente nivel de organización que es la estructura secundaria.

Existen dos principales conformaciones de estructura secundaria: la estructura hélice-α y la estructura lámina-β. Las interacciones que se forman entre estas estructuras secundarias (interacciones hidrofóbicas, puentes disulfuro, puentes de hidrógeno) permiten el plegamiento en una estructura tridimensional global conocida como estructura terciaria y cuando estas estructuras se encuentran compuestas por más de una cadena polipeptídica se conoce como estructura cuaternaria.

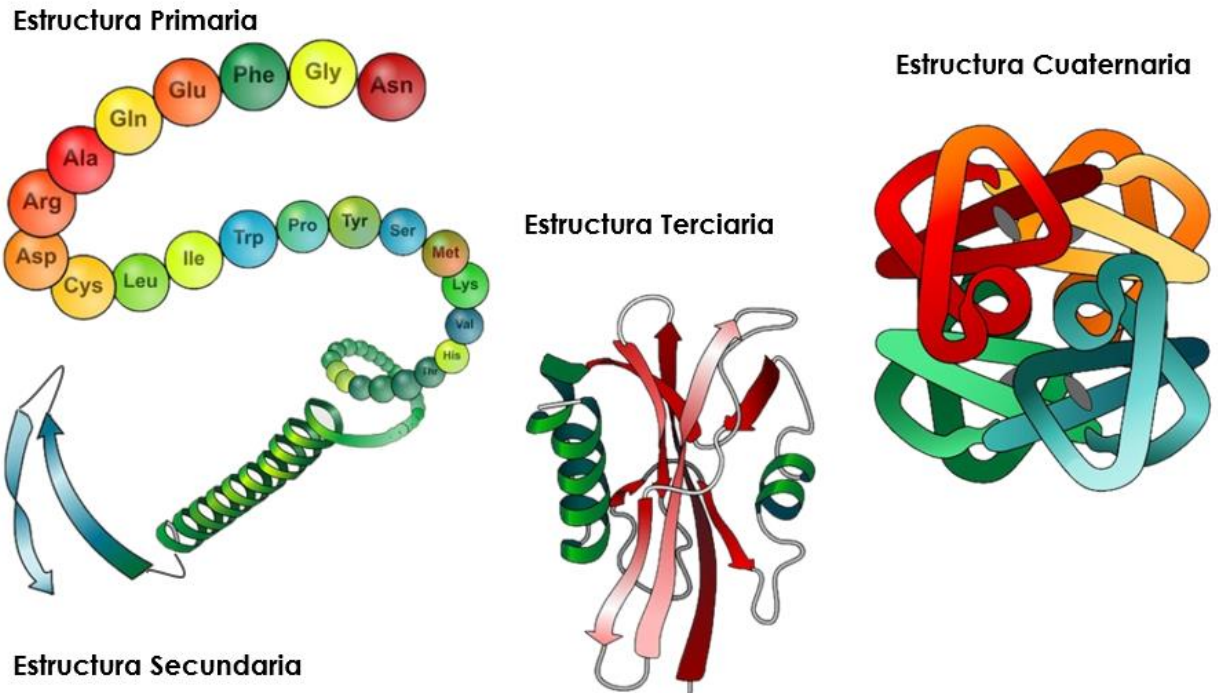


Figura 4. Niveles de organización en las proteínas. Modificado de Smith, Y. 2018. https://www.news-medical.net/life-sciences/Protein-Structure-and-Function-(Spanish).aspx

En la figura 4, se muestran los cuatro niveles de organización en proteínas, cada uno de estos niveles presenta gran diversidad, lo cual resulta en la gran variabilidad estructural observada en las proteínas la cual les permite realizar una amplia gama de funciones celulares.

Debido al mecanismo de "llave-cerradura" reconocido por Emil Fisher en 1894, en el estudio de proteínas se consideraba relevante el paradigma secuencia-estructura-función, el cual establece que la función de una proteína está determinada por su estructura, a su vez asociada a la secuencia de aminoácidos, permitiéndole interactuar con otras biomoléculas (ADN, ARN o proteínas); en otras palabras la enzima y el sustrato debían encajar entre sí como una cerradura y una llave para poder ejercer un efecto químico una sobre la otra. Posteriormente, en 1978 a través de técnicas de cristalografía de rayos X y de Resonancia magnética nuclear (NMR) se identificó el desorden estructural presente en las proteínas, el cual permitía cambiar de un estado conformacional a otro y por ende ejercer múltiples funciones celulares; desarrollándose la investigación en proteínas que presentaban desorden estructural posteriormente conocidas como proteínas intrínsecamente desordenada (IDPs) y la ruptura del paradigma estructura-función (Li et al., 2015; Tompa, 2012; Wright and Dyson, 1999).

2.2 Proteínas intrínsecamente desordenadas

Las proteínas o regiones/segmentos dentro de proteínas, caracterizadas por la falta de una estructura secundaria o terciaria estable en condiciones fisiológicas o en ausencia de ligando de unión Figura 5. Se denominan como proteínas intrínsecamente desordenadas (IDPs) o regiones intrínsecamente desordenadas (IDRs), estas proteínas presentan estructuras secundarias muy variables y no asumen las estructuras típicas como α-hélice o láminasβ (Ferron et al., 2006; Li et al., 2015; Wright and Dyson, 1999; Xue et al., 2010).

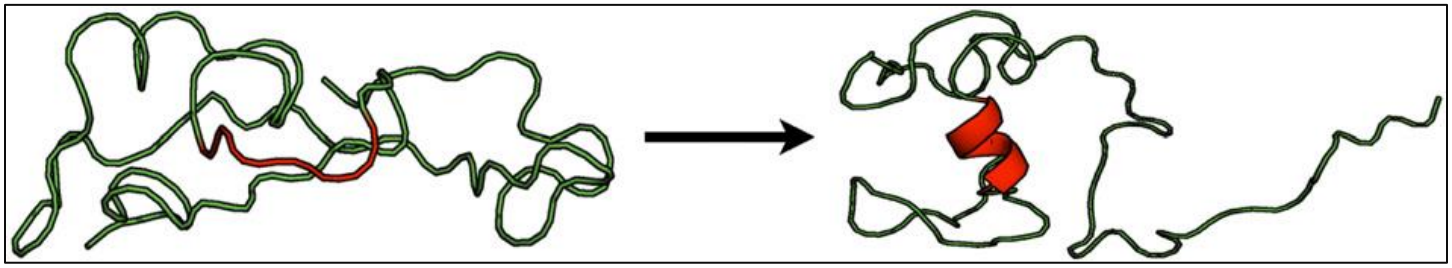


Figura 5. Imagen representativa de una proteína IDP tomado de Boskey and Villarreal-Ramírez, 2016.

El desorden estructural es común en los organismos eucariontes superiores y se estima que en humanos cerca de un tercio de todas las proteínas tiene un segmento intrínsecamente desordenado. Las IDPs desempeñan múltiples funciones, debido al desorden estructural presente que provee de múltiples ventajas funcionales importantes en la determinación de la respuesta celular, dentro de los cuales se incluyen reconocimiento, regulación y señalización (Alberts Bruce et al., 2016; Li et al., 2015). Actualmente, se sabe de la abundancia de las proteínas IDPs y de su amplia gama de funciones biológicas, así como de su relación con distintas enfermedades (Boskey and Villarreal-Ramírez, 2016).

2.3 Modificaciones postraduccionales.

Después de ser sintetizadas las proteínas son sujetas a muchos tipos de procesamiento post-traduccional en diferentes compartimientos celulares, mediante los cuales adquieren características moleculares transitorias o permanentes que les confieren varias propiedades tanto estructurales como funcionales. Estas modificaciones involucran procesamientos enzimáticos incluyendo la eliminación de uno o más aminoácidos del amino terminal, corte proteolítico o adición de grupos acetilo, metilo, glicosilo o fosfato a la cadena de aminoácidos y generan mecanismos dinámicos mediante los cuales las proteínas pueden cambiar de un estado a otro (Alberts Bruce et al., 2016).

2.3.1 Fosforilación

Bioquímicamente la fosforilación es la transferencia, catalizada enzimáticamente de un grupo fosfato de adenosin trifosfato (ATP) al grupo hidroxilo de la cadena lateral de un residuo aceptor, generando adenosin difosfato (ADP), en organismos eucariotas los aminoácidos capaces de ser fosforilados son: serina (S), treonina (T) y tirosina (Y), estos tres aminoácidos tienen la capacidad de unirse a los fosfatos porque todos ellos contienen un grupo hidroxilo en sus cadenas laterales que están desprotonadas a un pH fisiológico. Sin embargo, no todas las S, T y son susceptibles a la fosforilación, esta modificación es catalizada por un gran número proteínas miembros de la familia de las proteínas cinasas, que reconocen y fosforilan residuos de aminoácidos específicos y son la segunda familia de proteínas más grande en el genoma humano, esta gran familia de proteínas comparten dominios catalíticos homólogos y el mecanismo de reconocimiento de substrato puede ser similar (Alberts Bruce et al., 2016; Blom et al., 2004).

La fosforilación representa un importante mecanismo regulador en las células eucariotas, puesto que se estima que al menos un tercio de todas las proteínas eucariotas pueden fosforilarse afectando a la proteína de maneras importantes (Iakoucheva, 2004). En primer lugar el grupo fosfato aporta carga negativa que puede provocar un cambio conformacional importante, esto a su vez puede afectar la unión de ligando cambiando radicalmente la actividad de la proteína. En segundo lugar un grupo fosfato puede formar parte de una estructura que puede ser reconocida por otras proteínas (Alberts Bruce et al., 2016).

Esta modificación post-traduccional es reversible y un evento clave en el inicio de la transducción de señales de los sistemas bilógicos, regula la unión de los factores de transcripción a sus coactivadores y al DNA, lo que altera la expresión génica esencial para la regulación de procesos celulares como el metabolismo, proliferación, diferenciación y apoptosis (Iakoucheva, 2004).

Para comprender la actividad biológica de las proteínas, es necesario tener idea de cómo es su conformación espacial, por ende la importancia de saber el contenido de estructura secundaria, de estructura desordenada y de posibles modificaciones postraduccionales que de alguna manera participan en la adquisición de una determinada estructura y función.

Como se mencionó anteriormente existen diversas metodologías y técnicas computacionales para el análisis de secuencias de aminoácidos muchas de las cuales pertenecen a las máquinas de aprendizaje.
Una máquina de aprendizaje es un proceso adaptativo que permite a las computadoras aprender de la experiencia, aprender con el ejemplo y aprender por analogía. Las Redes Neuronales Artificiales (ANN) son máquinas de varios enfoques de aprendizaje, que se han aplicado con éxito a la solución de una amplia gama de problemas en el análisis bioinformáticos (Escobar et al., 2011).

3. Redes neuronales artificiales.

Las redes neuronales artificiales (ANN) pueden definirse como estructuras que comprenden elementos de procesamiento simple adaptativo densamente interconectados llamadas neuronas o nodos artificiales, que son capaces de realizar cálculos masivamente paralelos para el procesamiento de datos y la representación del conocimiento (Basheer and Hajmeer, 2000).

Las ANN presentan notables características de procesamiento de información del sistema biológico (Anil K. Jain et al., 1996), dentro de estas podemos mencionar:

- No linealidad
- Alto paralelismo
- Robustez
- Tolerancia a fallas y fallos
- Aprendizaje
- Capacidad de manejar información imprecisa y difusa
- Capacidad para generalizar

Los modelos artificiales que poseen tales características son deseables por las siguientes razones:

I. La no linealidad permite un mejor ajuste a los datos
II. La insensibilidad al ruido proporciona una predicción precisa en presencia de datos inciertos y errores de medición
III. Un alto paralelismo implica un procesamiento rápido y fallos de hardware la tolerancia
IV. El aprendizaje y la adaptabilidad permiten que el sistema actualice (modifique) su estructura interna en respuesta a un entorno cambiante
V. La generalización permite la aplicación del modelo a datos no aprendidos.

Las redes neuronales artificiales son un modelo computacional inspirado en su homólogo biológico; en las neuronas artificiales las conexiones entre nodos representan los axones y las dendritas, los pesos de conexión representan las sinapsis y el umbral se aproxima a la actividad en el soma (Anil K. Jain et al., 1996). Debido a que una neurona tiene un gran número de dendritas, puede recibir y transferir muchas señales simultáneamente, estas señales pueden inhibir o excitar la activación de otra neurona. Este mecanismo de transferencia de señal es el fundamento de la neurocomputación (computación basada en ANN).

La idea no es replicar el funcionamiento de los sistemas bilógicos, sino hacer uso de lo que se sabe sobre la funcionalidad de las redes biológicas para resolver problemas complejos, por lo que, el objetivo principal de la neurocomputación es desarrollar algoritmos matemáticos que permitan que las ANN aprendan imitando el procesamiento de información y la adquisición de conocimiento en el cerebro humano y actualmente han sido publicados muchos algoritmos de aprendizaje automático para el análisis y predicción de datos biológicos (Basheer and Hajmeer, 2000).

En la Figura 6 se muestran "n" neuronas biológicas con varias señales de intensidad "xy" fuerza sináptica "w" alimentando una neurona con un umbral de "b", y el sistema de neuronas artificiales equivalentes.

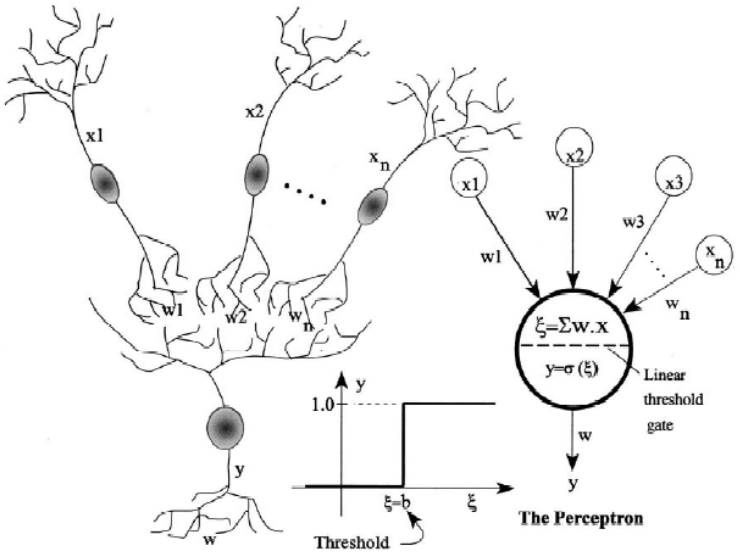


Figura 6. Interacción de las señales de "n" neuronas y la analogía con una neurona artificial que comprende un perceptrón de una capa tomado de Basheer and Hajmeer, 2000.

En este caso una neurona de procesamiento artificial recibe entradas como estímulos del entorno, las combina de una manera especial para formar una entrada "neta" (x), pasa a través de un umbral lineal y transmite la señal (salida, "y") hacia otra. La neurona o el medio ambiente, como se muestra en la Figura 6. Solo cuando la "x" excede el límite del umbral de la neurona ("b"), la neurona se activará. Comúnmente, se asume la dinámica de las neuronas lineales para calcular "x" (Haykin, 1994). La entrada "neta" se calcula como el producto interno (punto) de las señales de entrada (x) que inciden en la neurona y sus fuerzas (w). Para "n" señales, la operación de la neurona se expresa como:

$$y = \begin{cases} 1, si \ \sum_{i-1}^{n} w_i x_i \geq b, \\ 0, si \ \sum_{i-1}^{n} w_i x_i < b, \end{cases} \quad (2)$$

Donde 1 indica 'activación' y 0 indica 'inhibición'. Los pesos de conexión positivos ($w_i > 0$) mejoran la señal neta (x) y excitan la neurona, y el enlace se llama excitación, mientras que los pesos negativos reducen "x" e inhiben la actividad de la neurona, y el enlace se llama inhibitorio. La mecánica de la neurona artificial como se muestra en la figura 8 se llama Perceptron introducido por Rosenblatt en 1958.

El Perceptron establece un mapeo entre la actividad de las entradas (estímulos) y la señal de salida. En la ecuación, el umbral de la neurona se puede considerar como un nodo de entrada adicional cuyo valor es siempre la unidad (es decir, x =1) y su peso de conexión es igual a "b". En tal caso, la suma en la ecuación. (2) se ejecuta de 0 a "n", y la señal de red x se compara con 0.

Además de las diferencias en el método computacional usado, los predictores presentan diferentes combinaciones de características de entrada. Comúnmente utilizan características incluidas: la secuencia de aminoácidos, composición de aminoácidos, complejidad de la secuencia, matrices de puntaje de posición específica, predicción de estructura secundaria y la predicción de estructura accesible al solvente (Xue et al., 2010)

4. Herramientas para el análisis de secuencia de aminoácidos.

Cada tipo de proteína está formada por una secuencia precisa de aminoácidos que le permite plegarse en una forma tridimensional particular; existe una serie de factores que modifican esta estructura mencionados anteriormente; por esta razón las herramientas para el análisis se la secuencia de aminoácidos se dividieron en las siguientes categorías:

- Predicción del contenido de estructura secundaria
- Identificación de IDPs
- Predicción de posibles sitios de fosforilación

4.1 Predicción del contenido de estructura secundaria

La predicción de la estructura de proteínas a partir de la secuencia de aminoácidos es un reto importante para la biología molecular y un paso intermedio para la predicción de estructura terciaria puesto que es más fácil predecir aspectos simplificados de la estructura, es decir, los elementos estructurales clave de la proteína y la ubicación de estos elementos no en el espacio tridimensional sino a lo largo de la secuencia de aminoácidos (Kouza et al., 2017).

La determinación del contenido de estructura secundaria puede llevarse a cabo por técnicas experimentales como: Dicroísmo circular (CD), Espectroscopia infrarrojo y NMR (Cai et al., 2003); en los años 1970´s con el determinación de solo algunas estructuras tridimensionales mediante técnicas de difracción de rayos X y debido al alto costo de la técnica, se desarrollaron métodos de predicción computacionales que actualmente son rápidos, reproducibles, disponibles y lo suficientemente precisos por lo son una alternativa eficiente y de menor costo para la predicción de estructura secundaria (Cai et al., 2003; Chou and Fasman, 1977).

Considerando que el contenido de estructura secundaria es una característica importante tanto para estudios experimentales como teóricos, se han desarrollado diferentes métodos y algoritmos para su determinación, algunos de los cuales hacen uso de ANN (Cai et al., 2002).

4.1.1 CFSSP

Servidor de Predicción de Estructura Secundaria Chou & Fasman (Chou, Peter Y. and Fasman, Gereald D., 1974) es un predictor de estructura secundaria de proteínas a partir de la secuencia de aminoácidos implementado el algoritmo desarrollado por Chou-Fasman (Kumar, 2013).

Este algoritmo determina la tendencia de cada residuo a encontrarse en cada una de los tipos de estructura usando frecuencias observadas en estructuras proteínas resueltas con cristalografía de rayos X. El cálculo de la puntuación de propensión es simple; supongamos que hay "n" residuos en la estructura de la proteína de los cuales "m" son residuos en hélice, el número total de residuos de Alanina es "y" de los cuales "x" están en hélices. La puntuación de propensión para la Alanina de estar en una hélice está dada por la siguiente relación:

$$\frac{(\frac{x}{m})}{(\frac{y}{n})} \qquad (3)$$

Si la puntuación de propensión para un residuo es igual a 1.0 para hélice (P (hélice-a)) significa que el residuo tiene igual probabilidad de ser encontrado en una hélice o en cualquier otra estructura. Si P (hélice-a) < 1,0 entonces el residuo tiene poca oportunidad de ser encontrado en una hélice Si P (hélice-a) > 1,0 entonces es altamente probable que el residuo se encuentre en una hélice. Usando este concepto Chou y Fasman crearon la siguiente Tabla 3 (Chou, Peter Y. and Fasman, Gereald D., 1974).

Amino Ácido	(α-Hélice)	P (Lámina-β)	P (Giro)
Alanina	1.42	0.83	0.66
Arginina	0.98	0.93	0.95
Asparagina	0.67	0.89	1.56
Ácido Aspártico	1.01	0.54	1.46
Cisteína	0.70	1.19	1.19
Ácido Glutámico	1.51	0.37	0.74
Glutamina	1.11	1.11	0.98
Glicina	0.57	0.75	1.56
Histidina	1.00	0.87	0.95
Isoleucina	1.08	1.60	0.47
Leucina	1.21	1.30	0.59
Lisina	1.14	0.74	1.01
Metionina	1.45	1.05	0.60
Fenilalanina	1.13	1.38	0.60
Prolina	0.57	0.55	1.52
Serina	0.77	0.75	1.43
Treonina	0.83	1.19	0.96
Triptófano	0.83	1.19	0.96
Tirosina	0.69	1.47	1.14
Valina	1.06	1.70	0.50

Tabla 3. Puntuación de propensión para estructura secundaria tomado de Chou and Fasman, 1974.

El algoritmo Chou-Fasman toma la secuencia y la divide en ventanas de tamaño fijo para determinar el número de residuos pertenecientes a cada estructura usando la puntuación de propensión. Para hélices-a la ventana es de tamaño 6, si una región tiene 4 residuos contiguos cada uno con P (hélice-a) > 1,0, se concluye que el conjunto forma parte de una hélice, ésta región en hélice se extiende en ambas direcciones hasta que P (hélice-a) < 1,0. Para láminas-β utiliza una ventana de 5 residuos, si se tienen al menos 3 residuos cada uno con P (lámina-β) > 1,0, se concluye que el conjunto forma parte de una lámina-β, si ambos tipos de estructuras se traslapan en cierta región, se toma la siguiente decisión Si Σ P (hélice-a) > Σ P (lámina-β) entonces se concluye una hélice-a (Kumar, 2013).

Como resultado el servidor genera la predicción de estructura secundaria se muestra en una vista gráfica lineal secuencial basada en la probabilidad de ocurrencia de estructura α-hélice, lámina-β y giros.

4.1.2 GOR IV

El método original fue publicado por Garnier, Osguthorpe y Robson en 1978 (Kouza et al., 2017) es un método basado en la teoría de la información para la predicción de estructuras secundaria en proteínas a partir de la secuencia de aminoácidos (Garnier, Jean et al., 1996). El método GOR se basa en parámetros de probabilidad derivados de estudios empíricos de estructuras terciarias de proteínas conocidas resueltas mediante cristalografía de rayos X. Sin embargo, a diferencia de Chou-Fasman, el método GOR tiene en cuenta no solo las propensiones de los aminoácidos individuales para formar estructuras secundarias particulares, sino también la probabilidad condicional de que el aminoácido forme una estructura secundaria dado que sus vecinos inmediatos ya han formado esa estructura (Garnier, Jean et al., 1996).

El método se basa en la teoría de la información y en el supuesto de que la información de la función de una proteína se puede predecir mediante una suma de información de residuos individuales y pares de residuos. Una de las herramientas matemáticas básicas de la teoría de información es la función de información I (S, R):

$$I(S;) = log[P(S|R)/P(S)] \quad (4)$$

Para el problema de predicción de estructura secundaria de proteínas, la función de información se define como el algoritmo de la relación de probabilidad condicional P (S | R) de observar la conformación S, donde S es uno de los posibles estados conformacionales para el residuo R, donde R es uno de los 20 aminoácidos posibles y la probabilidad P(S) de la aparición de la conformación S. La función de información I (S; R) es calculado a partir de una base de datos de proteínas utilizada en el programa (267 proteínas para GOR IV).

El estado conformacional de un residuo dado en la secuencia depende no solo del tipo de aminoácido R sino también de los residuos vecinos a lo largo de la cadena dentro de la ventana de búsqueda (GOR IV usa una ventana de 17 residuos, es decir, para un residuo dado, se analizaron ocho residuos vecinos más cercanos a cada lado). La ventaja del método GOR sobre otros métodos es que identifica claramente todos los factores que se incluyen en el análisis y calcula las probabilidades de posibles estados conformacionales.

El programa analiza las secuencias para calcular la puntuación de propensión para cada residuo de estar en cada una de las cuatro posibles estructuras: α-hélice (H), lámina-β (E), giros (T) o aleatoria (C), en función de las interacciones con los residuos vecinos, por lo cual examina una ventana de 17 aminoácidos y suma la propensión de las cuatro posibles estructuras (cuatro sumatorias), la estructura con la puntuación más alta define el tipo de estructura al que pertenece el residuo al centro de la ventana (noveno residuo).

En la interface del programa se solicita como datos de entrada: el nombre de la secuencia (opcional), la dirección de correo electrónico del usuario (al cual serán enviados los resultados de la predicción) y la secuencia de la proteína de interés en código de una sola letra (el servidor acepta un mínimo de 20 y máximo de 1000 aminoácidos). Como resultado presenta la predicción de estructura secundaria de la secuencia de entrada y las probabilidades para cada elemento de estado secundario en cada posición, el porcentaje final para cada uno de los tipos de estructura, así como una representación gráfica de los resultados.

4.1.3 SOPMA

El método de predicción auto-optimizado con alineación (SOPMA) es una mejora del método SOPM (Geourjon and Deléage, 1994). Estos métodos basados en homología, no sólo toman en cuenta la estadística, también consideran patrones comunes conservados entre múltiples secuencias homólogas; emplean información evolutiva, combinan métodos ab initio para predicción de la estructura secundaria de secuencias individuales e información de alineamiento múltiple de secuencias homólogas. La idea detrás de este enfoque es que proteínas homólogas adoptan la misma estructura secundaria y terciaria (Geourjon and Deléage, 1995).

4.2 Identificación de IDPs

Las funciones biológicas significativas de las IDPs, han desencadenado el desarrollo de metodologías para la identificación de las mismas, incluyendo metodologías como: la digestión con proteasas, difracción de rayos X, NMR, DC y diversas mediciones hidrodinámicas (Ferron et al., 2006; Romero et al., 2001). Otro enfoque para la identificación de IDPs ha sido a través de métodos computacionales utilizando predictores del estado de desorden, en los cuales se analizan las secuencias de aminoácidos codificadas por los genes y se buscan las regiones desordenadas con base a las características de los aminoácidos (Alberts Bruce et al., 2016) este tipo de metodología ha sido importante por las siguientes razones (Ferron et al., 2006; Li et al., 2015).

I. Identificar regiones desordenadas puede promover el análisis de proteínas.
   Las regiones desordenadas en una proteína tienen composición de aminoácidos sesgada (presenta repeticiones cortas y mayor presencia de unos pocos residuos) que puede dar lugar a alineamientos de secuencias inexactos a proteínas no relacionadas. Al reconocer las regiones desordenadas, se puede evitar alinear estas regiones con las regiones ordenadas y, por lo tanto, aumentar la precisión del análisis de similitud de secuencias.
II. Las regiones desordenadas dificultan la purificación y cristalización de una proteína.
   La identificación de proteínas como altamente desordenadas puede ahorrar tiempo, ya que no se gastaría tiempo intentando determinar una estructura que no existe.
III. Identificación de motivos lineales eucariotas
   El reconocimiento de regiones desordenadas facilita la identificación de motivos lineales eucariotas (ELM), que son motivos funcionales cortos que se presentan principalmente (>70%) dentro de regiones desordenadas.

Las IDPs tienen una **composición de aminoácidos sesgada**, de manera que se ha establecido la siguiente regla empírica: glicina, serina y prolina son aminoácidos que promueven el desorden, triptófano, fenilalanina, isoleucina, tirosina, valina y leucina son aminoácidos que promueven el orden, mientras que histidina y treonina se consideran neutrales con respecto al desorden**; la predicción de contenido de estructura secundaria es bajo,** ya que cuentan con una composición de aminoácidos sesgada hacen uso de menos tipos de aminoácidos por lo que tienden a tener una **complejidad de secuencia baja** y las regiones desordenadas presentan **alta variabilidad de secuencia** en comparación con las regiones ordenadas. (Ferron et al., 2006).

Se han desarrollado varios algoritmos para predecir regiones desordenadas utilizando las características se secuencias desordenadas antes mencionadas. Estos métodos computacionales son capaces de producir una puntuación de predicción de residuos desordenados en IDPs, proporcionando una solución razonable para el ahorro de tiempo y el costo experimental.

Existen diferentes clasificaciones para los predictores. Según Ferron et al., 2006 se pueden clasificar en dos tipos: aquellos que han sido entrenados en bases de datos de proteínas desordenadas y aquellos que no; los predictores que se basan en el entrenamiento contra una base de datos de regiones de proteínas desordenadas presentan algunas inconsistencias ya que las bases de datos contienen relativamente pocas proteínas desordenadas.

Según Li et al., 2015, los predictores pueden clasificarse como:

- Predictores basados en clasificadores de aprendizaje automático
- Predictores basados en un meta-enfoque que combina predicciones de múltiples predictores
- Predictores basados en las propiedades fisicoquímicas.

La predicción de proteínas desordenadas puede manejarse como un problema de clasificación binaria (ordenado/desordenado) y se puede dirigir a métodos de máquinas de aprendizaje automático, como las redes neuronales artificiales (ANN).

4.2.1 PONDR

Es un predictor que utiliza una red neuronal basada en la composición local de aminoácidos, flexibilidad y otras características de secuencia; fue el primer predictor que se desarrolló y actualmente se encuentra disponible en varias versiones (Ferron et al., 2006).

Las extensiones agregadas a PONDR describen los datos de entrenamiento de un predictor en particular. La primera letra se refiere al método de caracterización:

- X para Rayos X
- N para NMR
- C para dicroísmo circular
- V para varios

La segunda letra hace referencia a la longitud o ubicación de la región:

- S para 8-9 residuos
- M para 20-39 residuos
- L para 40 o más residuos
- N para terminal amino 5 o más residuos
- C para terminal carboxilo 5 o más residuos
- T residuos en cualquier termino 5 o más residuos

4.2.2 PONDR VL-TX

Este predictor utiliza un clasificador de red neuronal no lineal, entrenado para distinguir desordenado / ordenado según características como la composición de los aminoácidos, complejidad de la secuencia, la carga neta, la hidropatía promedio en ventanas de 9 a 21 aminoácidos; como "input" se introduce el código NCBI o la secuencia de aminoácidos en formato FASTA.

Este predictor aplica tres diferentes redes neuronales, una por cada región terminal y una para la región interna de la secuencia. Cada red neuronal está entrenada por un conjunto de datos específico que contiene solo los residuos de aminoácidos de esa región específica. El resultado de la predicción final utiliza los predictores individuales en sus respectivas regiones (Xue et al., 2010)

- VL1: entrenado utilizando 8 regiones largas desordenadas identificadas por densidad de electrones faltantes en estudios cristalográficos de rayos X, y 7 regiones largas caracterizadas por RNM (Romero et al., 1997).
- TX: Son dos predictores, XN: predictor N-terminal y XC: predictor C-terminal (Li, Xiaohong et al., 1999) se entrenaron utilizando datos cristalográficos de rayos X, donde las regiones terminales tenían 5 o más residuos de longitud.

Los "outputs" para el predictor VL-TX consisten en tres partes:

- Una representación gráfica
- Una salida textual para demostrar el desorden a lo largo de la secuencia
- Puntaje bruto para cada aminoácido de la proteína: Son números, donde 1 es la predicción ideal de desorden y 0 la predicción ideal de orden, generalmente los resultados de predicción no son ideales entonces, si el valor supera o coincide con un umbral de 0.5 el residuo se considera desordenado.

4.3 Predicción de potenciales sitios de fosforilación

Debido a la relevancia biológica de esta modificación se han desarrollado muchas y diversas técnicas experimentales que permiten detectar la fosforilación, como son: inmunoprecipitación con anticuerpos, técnicas de cromatografía de afinidad basadas en el uso de metales en micro-columna, espectrometría de masas (MS) y otras metodologías basadas en el uso de máquinas de aprendizaje automático para la predicción de sitios de fosforilación (Dephoure et al., 2013; López Villar et al., 2008).

La identificación del mecanismo de reconocimiento de sustrato de las proteínas cinasas ha sido un gran reto para en la investigación. Se ha descubierto mediante estudios de cristalización que la región de entre siete y doce residuos que rodean al residuo aceptor contactan con el sitio activo de la cinasa (Blom et al., 2004), por ello que la secuencia primaria que flanquea al residuo fosfoaceptor juega un papel importante en definir el potencial del sustrato de ser fosforilado. Sin embargo, diversos estudios han revelado que la secuencia de aminoácidos locales no es determinante y que otros factores incluyendo la estructura espacial local y la accesibilidad de la superficie también son importantes (Kreegipuu et al., 1999).

Basados en sets de sitios de fosforilación verificados experimentalmente, Blom et al., generaron las secuencias logotipo para cada uno de los tres aminoácidos aceptores, analizaron las secuencias locales que rodean a los residuos fosfoaceptores de fosfoserina, fosfotreonina y fosfotirosina para poder identificar qué características presentan estos residuos que forman parte de los motivos de reconocimiento de las proteínas cinasas e identificar estos patrones, determinaron que la especificidad de las proteínas cinasas está sujeta a la presencia de residuos ácidos, básicos o hidrofóbicos adyacentes (Blom et al., 1999).

Debido a la gran variación de los residuos que pueden ocupar estos sitios adyacentes al residuo fosfoaceptor, se dificulta el realizar una búsqueda manual de las secuencias de las proteínas y predecir la posición biológica de sitios activos, por esta razón que el uso de ANN para la predicción del estado fosforilado ha sido importante, ya que son capaces de clasificar grandes complejos de patrones de secuencias biológicas donde la correlación entre posiciones es importante.

Existen varios algoritmos que han sido implementados en estrategias de predicción, desde simples buscadores de motivos hasta los más complejos métodos como las ANN, donde la red reconoce los patrones vistos durante el entrenamiento (ocupa la información de las bases de datos) y mantiene la habilidad de generalizar y reconocer patrones similares, pero no idénticos para la discriminación entre posibles sitios de fosforilación y los cuales aparentemente nunca estarían modificados (Blom et al., 2004, 1999).

4.3.1 NetPhos

NetPhos es un programa que predice el estado fosforilado de proteínas a partir de la secuencia de aminoácidos en nomenclatura de una sola letra o en formato FASTA. Los resultados se basan en comparaciones de secuencia y estructura con sitios de fosforilación verificados de la base de datos de PhosphoBase (Kreegipuu et al., 1999). Esta base de datos es una compilación de la información acerca de los sitios de fosforilación de proteínas reportadas en la literatura.

El contenido de la base de datos consiste en 414 fosfoproteínas, hay proteínas de 67 diferentes organismos la mayoría especies de vertebrados (humano, rata, bovino, ratón, conejo y gallina), pero también fosfoproteínas de plantas, insectos, bacterias y virus; todas estas proteínas en conjunto contienen un total 1052 sitios de fosforilación determinados experimentalmente, entre los cuales hay 688 serinas, 168 treoninas y 196 tirosinas fosforilables por cerca de 100 diferentes proteínas cinasas; por lo que esta base de datos provee de buenas bases para el análisis de secuencias substrato preferidas por las proteínas cinasas (Kreegipuu et al., 1999).

NetPhos utiliza una ANN para predecir los sitios de fosforilación en proteínas eucariotas, la red neuronal esta entrenada para el reconocimiento de patrones; ya que el contenido de la base de datos consiste completamente en secuencias; para cualquier proteína fosforilada, la base de datos proporciona la secuencia completa de las 414 fosfoproteínas, seguida de observaciones sobre qué residuos están fosforilados en el contexto de +/- 4 residuos que flanquean al aceptor de fosfato. Por lo tanto NetPhos tiene la capacidad de identificar nuevos sitios de fosforilación mediante la identificación de las características prevalentes comunes en las secuencias que rodean al residuo fosfoaceptor de fosfoproteínas conocidas.

En el programa se pueden realizar predicciones genéricas o específicas de cinasas donde las predicciones se hacen para las siguientes 17 proteínas: ATM , CKI , CKII , CaM-II , DNAPK , EGFR , GSK3 , INSR , PKA , PKB , PKC , PKG , RSK , SRC , cdc2 , cdk5 y p38MAPK .

Después de que se ingresa la secuencia de aminoácidos en la interfaz de NetPhos, el programa genera una tabla que incluye: el número de residuo, el tipo y contexto en el que se encuentra el aminoácido, así como una puntuación que va de 0 a 1 (donde el 0 es el valor correspondiente a una fosforilación negativa) y la proteína cinasa asociada.

5. YASARA

YASARA (Krieger and Vriend, 2014). Es un paquete de programas de computadora para aplicaciones de modelado molecular y dinámica molecular, se encuentra disponible en cuatro paquetes diferentes: YASARA View, adecuado para análisis y gráficos moleculares; para modelado molecular y dinámica molecular existen YASARA Model y YASARA Dynamics; y por ultimo YASARA Structure, para el modelado de homología y simulaciones de acoplamiento molecular. El modelado molecular es la técnica general para modificar y simular una estructura de proteína *in silico* mediante técnicas computacionales; este programa permite realizar modificaciones como la fosforilación en los modelos de proteínas o péptidos (Land and Humble, 2018).

6. PyMOL

PyMOL (Delano, W.L., 2002). Es un programa de gráficos moleculares apropiado para la visualización de múltiples configuraciones de una sola estructura (trayectorias) y la generación de imágenes 3D de alta calidad de macromoléculas biológicas, como las proteínas. La visualización está disponible con diferentes representaciones para estructuras macromoleculares que se pueden cargar desde archivos PDB, así como varios otros formatos de archivo. Además de ser posible cargar trayectorias y conjuntos conformacionales directamente en PyMOL para la visualización dinámica se pueden generar una secuencia de imágenes de estados moleculares como una serie de archivos numerados para ensamblar películas en QuickTime o AVI.

Simulaciones de dinámica molecular

La simulaciones de dinámica molecular (SDM) es uno de los métodos alternativos para la realización de simulación computacional, es un método a nivel atomístico que permite analizar el comportamiento o evolución de un sistema (físico, químico o biológico) a través del tiempo, calculando las fuerzas entre los átomos que lo conforman mediante las ecuaciones de movimiento de Newton (Lozano-Aponte and Scior, 2014).

Macromoléculas biológicas como las proteínas, son sistemas intrínsecamente dinámicos y los movimientos macromoleculares cubren un amplio rango de magnitud (desde centésimas de un angstrom a decenas de angstroms) y un enorme intervalo de tiempo (desde sub-picosegundos a segundos e incluso más). Se sabe que algunos de estos movimientos tienen un papel funcional, por ello que las simulaciones de dinámica molecular son importantes para comprender estos movimientos (Karplus and Petsko, 1990).

Las simulaciones de dinámica molecular comienzan con un conocimiento de la energía del sistema en función de las coordenadas atómicas. Las fuerzas que actúan sobre los átomos del sistema, que están relacionadas con las primeras derivadas del potencial con respecto a las posiciones de los átomos, pueden usarse para calcular el comportamiento dinámico del sistema resolviendo las ecuaciones de movimiento de Newton para los átomos como una función de tiempo.

Las funciones de energía utilizadas para las proteínas generalmente están compuestas por términos de enlace que representan longitudes de enlace, ángulos de enlace y ángulos de torsión, y términos de no enlace que consisten en interacciones de van der Waals y contribuciones electrostáticas. Una expresión muy utilizada es:

$$E(R) = \frac{1}{2}\sum_{bonds} K_b\,(b - b_0) + \frac{1}{2}\sum_{bond\ angles} K_\theta\ (\theta - \theta_0)^2$$

$$+ \frac{1}{2}\sum_{torsional} K_\emptyset\,[\ddot{}1 + \cos\,(n\emptyset - \delta)]$$

$$+ \frac{1}{2}\sum_{nb\ pairs}\left(\frac{A}{r^{12}} - \frac{B}{r^6} + \frac{q^1 q^2}{Dr}\right)$$

(5)

En la ecuación (5) la energía, E, es una función del conjunto de coordenadas cartesianas, R, que especifica las posiciones de todos los átomos, a partir de los cuales se calculan las coordenadas internas para las longitudes de enlace (b), ángulos de enlace (θ), ángulos diedros ($\emptyset$) y distancias entre partículas (r).

El primer término en la ecuación (5) representa desplazamientos instantáneos de la longitud de enlace ideal, b_0, por el potencial de la ley de Hooke (armónico). Tal potencial armónico es la primera aproximación a la energía de un enlace en función de su longitud. La constante de fuerza de enlace K_b determina la flexibilidad del enlace y puede evaluarse a partir de frecuencias de estiramiento infrarrojo o cálculos de mecánica cuántica. Las longitudes de enlace ideales se pueden inferir a partir de estructuras de alta resolución y baja temperatura o datos de espectroscopia de microondas. La energía asociada con la alteración de los ángulos de enlace dada por el segundo término en la ecuación (5) también se representa por un potencial armónico. Para las rotaciones sobre los enlaces, se utilizan las funciones de potencial de ángulo de torsión dadas por el tercer término en la ecuación (5). Se asume que este potencial es periódico y está modelado por un coseno o suma sobre las funciones del coseno. El término final en la ecuación (5) representa la contribución de las interacciones no vinculadas y tiene tres partes: un término repulsivo que evita que los átomos se intercalen a distancias muy cortas; un término atractivo que explica las fuerzas de dispersión de London entre los átomos; y un término electrostático que es atractivo o repulsivo dependiendo de si las cargas q^1 y q^2 tienen el signo opuesto o el mismo. Los dos primeros términos no vinculados se combinan para dar el potencial familiar de Lennard-Jones, que tiene un mínimo en una separación interatómica igual a la suma de los radios de van der

Waals de los átomos; los parámetros A y B dependen de los átomos involucrados y han sido determinados por una variedad de métodos, que incluyen distancias de no unión en los cristales y mediciones de dispersión de la fase gaseosa (Karplus and Petsko, 1990).

Las interacciones electrostáticas entre pares de átomos están representadas por un potencial de Coulomb con D, la función dieléctrica efectiva para el medio y r la distancia entre las dos cargas. El uso de cargas parciales atómicas evita la necesidad de un término separado para representar la interacción del enlace de hidrógeno; es decir, cuando el hidrógeno positivo unido a un átomo electronegativo llega a la distancia de van der Waals de un átomo aceptor negativo, la atracción de Coulomb aumenta el potencial de Lennard-Jones y da como resultado un enlace de hidrógeno (Karplus and Petsko, 1990).

La utilidad de las funciones de energía empírica depende de la medida en que los parámetros determinados para la ecuación (5) mediante el estudio de sistemas modelo, como los aminoácidos, pueden emplearse para macromoléculas, como las proteínas.

Dada una función de energía potencial, las ecuaciones de movimiento de Newton se resuelven para los átomos del sistema y cualquier solvente circundante. Para un sistema simple y homogéneo, como una caja de moléculas de agua con condiciones de frontera periódicas, se pueden determinar propiedades estructurales y dinámicas promedio en simulaciones de solo unos pocos picosegundos. Los sistemas no homogéneos, como las proteínas, requieren simulaciones considerablemente más largas de hasta nanosegundos.

Para comenzar una SDM, se requiere un conjunto inicial de coordenadas atómicas y velocidades:

Las coordenadas se pueden obtener a partir de los datos de la estructura cristalográfica de rayos X o de NMR, o mediante la construcción de modelos. Dado un conjunto de coordenadas, un cálculo preliminar sirve para equilibrar el sistema. La estructura se refina primero utilizando un algoritmo de minimización para aliviar las tensiones locales debidas a la superposición de átomos no unidos, distorsiones de longitud de enlace, etc.

A continuación, a los átomos se les asignan velocidades (v) tomadas al azar de una distribución maxwelliana para una temperatura baja, y se realiza una simulación durante unos pocos picosegundos. Esto se hace encontrando la aceleración a_i; del átomo i de la ley de Newton $F_i = m_i a_i$ (F_i, la fuerza en el átomo, se calcula a partir de las derivadas de la ecuación (5) con respecto a la posición; m_i, es la masa atómica), e introduciéndola en la ecuación para la posición r; en el tiempo $t + \Delta t$, dado r_i; en el momento t:

$$r_i\,(t + \Delta t) = r_i\,(t) + v_i\,\Delta t + \frac{1}{2} a_i\,(\Delta t)^2 + \cdots$$

(6)

El equilibrio se continúa alternando nuevas asignaciones de velocidad, elegidas de las distribuciones maxwellianas para temperaturas que aumentan sucesivamente a algún valor elegido, con intervalos de relajación dinámica. La temperatura T del sistema se mide por la energía cinética media:

$$\frac{1}{2}\sum_{i=1}^{N} m_1\,\langle v_i^2\rangle = \frac{3}{2}\,N k_b\,T$$

(7)

Donde N es el número de átomos del sistema, $\langle v_i^2\rangle$ es la velocidad media al cuadrado del átomo i y k_b es la constante de Boltzmann. El período de equilibrio se considera terminado cuando la temperatura es estable durante más de aproximadamente 10 ps, los momentos atómicos obedecen a una distribución maxwelliana y diferentes regiones de la proteína tienen la misma temperatura promedio.

La integración continua de las ecuaciones de movimiento después del equilibrio genera las coordenadas y velocidades de los átomos en función del tiempo. La ecuación (6) se puede resolver simultáneamente para todos los átomos en una macromolécula solvatada mediante el uso de una serie de programas estándar. La cantidad (Δt) debe ser muy pequeña para que la energía potencial no cambie demasiado durante cada paso del tiempo; un valor satisfactorio para (Δt) es un femtosegundo (10^{-15} s), que corresponde a unos 30 pasos para una vibración de un enlace carbono-carbono.

Para sistemas solvatados como es el caso de las proteínas y los ácidos nucleicos que funcionan en un ambiente acuoso; un gran número de moléculas de agua están estrechamente ligadas a sus superficies, formando una primera capa de hidratación que es parte integral de la estructura y debe incluirse en la simulación. Eso puede agregar cientos de átomos a la computación, y si bien el solvente a granel es un problema debido a que el tiempo aumenta de forma aproximadamente lineal con el número de átomos en el sistema, es aún más grave si no se incluye, la simulación 'en vacío', puede no afectar a los átomos interiores pero es probable que tenga un efecto grave en las conformaciones y movimientos de los residuos de la superficie, especialmente en los bucles expuestos. El disolvente debe incluirse en las simulaciones relacionadas con el análisis cuantitativo de las propiedades termodinámicas, particularmente de aquellas porciones de la proteína en la interfaz del disolvente.

Si bien se ha dependido de los resultados experimentales para poder parametrizar y validar los modelos de simulación molecular, lo que implica que la simulación ha seguido a la experimentación, se espera que dichos roles sean intercambiados, es decir que la simulación ayude a predecir resultados experimentales, así reducir la cantidad de recursos que se necesitan para la experimentación de manera confiable.

- Equilibrio: Se resuelven las ecuaciones de movimiento del sistema hasta que las propiedades del sistema no cambian más con el tiempo, como la energía total.
- Producción: son las mediaciones que se utilizan para los cálculos computacionales, puesto que el sistema ya se encuentra en equilibrio.

1. GROMACS

GROMACS es un acrónimo para GROningen MAchine for Chemical Simulation; es un paquete de programas de SDM, especialmente dirigido a la simulación de moléculas biológicas en entornos acuosos y de membrana; incluye varios métodos de acoplamiento de temperatura y presión y los átomos pueden organizarse en grupos especiales con el fin de la participación selectiva en la dinámica o el análisis detallado de las energías. Además GROMACS incluye una gran variedad de herramientas de análisis, que abarcan desde análisis de trayectorias gráficas extensas hasta modos normales y análisis de componentes principales de fluctuaciones estructurales (Mark Abraham et al., 2014).

- Definición del sistema.

El sistema se define por su tamaño y forma, el número y los tipos de moléculas que contiene, y las coordenadas y velocidades de todos los átomos. La geometría habitual para el sistema simulado es una celda rectangular con condiciones de contorno periódicas. GROMACS permite una forma general de celda triclínica, que abarca todas las posibles construcciones de espacio, además también es posible estudiar un sistema aislado, ya sea como tal o eligiendo una caja suficientemente grande (Van Der Spoel et al., 2005).

Los átomos del sistema a simular se colocan en una celda unitaria, rodeada por copias de sí misma; es decir en condiciones periódicas de frontera, como se muestra en la Figura 7 (Gustav-Stresemann et al., 2004)en el cual el sistema se encuentra en una celda central y esta es replicada de manera infinita hacia dos dimensiones (2D).

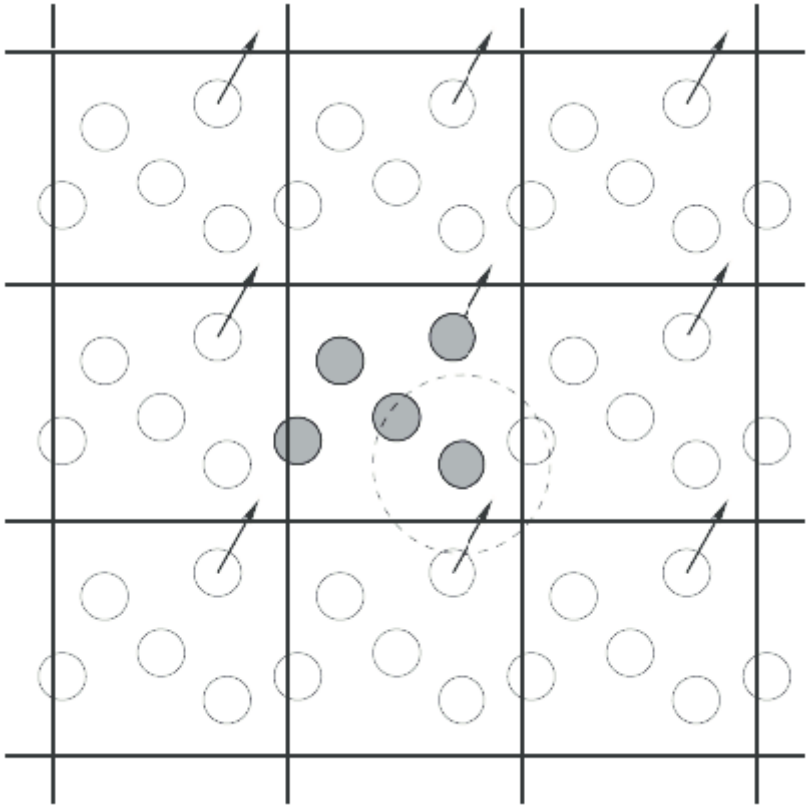


Figura 7. Representación de condiciones periódicas de frontera. Tomado de Gustav-Stresemann 2004.

- Campo de fuerza

Un campo de fuerza es una expresión matemática que describe la forma funcional y parámetros utilizados para calcular la energía potencial de un sistema de átomos o partículas en SDM. Proporciona parámetros para cada tipo de átomo en el sistema y los parámetros de las funciones de energía pueden derivarse de experimentos en física o química (difracción de rayos X, difracción de electrones, NMR, Infrarrojo, Raman, etc.) cálculos en mecánica cuántica (ad initio o semi empíricos) o ambos (González, 2011).

La forma funcional básica de energía potencial en mecánica molecular incluye términos enlace para interacciones de átomos que están vinculados por enlaces covalentes, y términos de no enlace (también denominados no covalentes) que describen las fuerzas electrostáticas y van der Waals de largo alcance. Una forma general para la energía total en un campo de fuerza aditiva se puede escribir como:

$$E_{total} = E_{covalentes} + E_{no\ covalentes} \tag{8}$$

Donde los componentes de las contribuciones enlace y no enlace están dados por las siguientes sumas:

$$E_{covalentes} = E_{enlace} + E_{ángulo} + E_{diédricos}$$
$$E_{no\ covalentes} = E_{electrostática} + E_{van\ der\ Waals} \tag{9}$$

Básicamente, los átomos se tratan como masas puntuales y la energía potencial del sistema de divide en varias contribuciones:

$$E_{total} = E_r + E_\theta + E_\emptyset + E_q + E_{vdw} + E_{acop}$$

- E_r, es la energía de tensión de los enlaces. Se considera cada enlace como un resorte con una distancia de equilibrio y una constante de elasticidad, cuyos valores dependen del tipo de enlace del que se trate.
- E_θ, es la energía de flexión. Dos enlaces con un átomo en común forman un ángulo, que tiene el valor de equilibrio y una constante de rigidez.
- $E_\emptyset$, es la energía de torsión. Cada ángulo define un plano, los planos formados por dos ángulos que comparten un enlace forman un ángulo de torsión.

- E_q, es la energía de la interacción electrostática. A cada átomo se le asigna una carga q y la interacción total se obtiene.
- E_{vdw}, es la energía de van der Waals, que incluye la repulsión electrostática y las interacciones de dispersión entre los distintos átomos. Generalmente se modela mediante un potencial de Lennard-Jones.
- E_{acop}, es la energía de acoplamiento entre los términos anteriores, que viene dada por términos cruzados que implican distintas distancias o ángulos.

Los campos de fuerza definen un conjunto de parámetros para diferentes tipos de átomos, enlaces químicos, ángulos diédricos, etc. El conjunto de parámetros incluye valores para la masa atómica, el radio de van der Waals y la carga parcial para los átomos individuales, y los valores de equilibrio de las longitudes de enlace, los ángulos de enlace y los ángulos diédricos para pares, tripletes y cuádruples de los átomos enlazados, y los valores correspondientes a la constante de resorte (Ley de Hooke) efectiva para cada potencial.

- Minimización de energía

En muchos casos se requiere de minimización energética; normalmente las estructuras iniciales que se construyen poseen energías muchos mayores a las que tendría un objeto real, por esta razón, se utilizan algoritmos para calcular las posiciones y fuerzas originales, con el objetivo de minimizarlas y que sean más realistas. Si una configuración inicial está muy lejos del equilibrio, las fuerzas pueden ser excesivamente grandes y la simulación de SDM puede fallar. Otra razón para realizar una minimización de energía es la eliminación de toda la energía cinética del sistema: si se deben comparar varias "imágenes" de simulaciones dinámicas, la minimización de energía reduce el ruido térmico en las estructuras y las energías potenciales para poder compararlas mejor (Mark Abraham et al., 2014).

- Ensamblaje NVT

Mientras que el uso directo de la SDM da lugar al ensamble NVE (número constante de partículas, volumen constante, conjunto de energía constante), la mayoría de las cantidades que deseamos calcular son en realidad un ensamble a temperatura constante NVT (número constante de partículas, volumen y temperatura), también llamado ensamble canónico (Mark Abraham et al., 2014) En NVT la energía de los procesos endotérmicos y exotérmicos se intercambia con un termostato, los algoritmos de termostato tienen como propósito generar un ensamble termodinámico a temperatura constante, lo cual se logra con la modificación del esquema Newtoniano de dinámica molecular. Existe una variedad de algoritmos de termostato disponibles incluidos los termostatos de Nosé-Hoover, Berendsen, Andersen por mencionar algunos (González, 2011; Mark Abraham et al., 2014).

- Ensamblaje NPT

Los enfoques para controlar la presión son similares a los empleados para la temperatura, en el ensamblaje isotérmico- isobárico NPT (número constante de partículas, presión y temperatura) se requiere además de un termostato un baróstato.

GROMACS resuelve las ecuaciones de movimiento de Newton para un sistema de N átomos que interactúan:

$$m_i \frac{\partial^2 r_i}{\partial t^2} = F_{i,}\ i = 1 \dots N \tag{10}$$

Las fuerzas son las derivadas negativas de una función potencial V ($r_1, r_2 \dots r_N$):

$$F_i = \frac{\partial V}{\partial r_i} \tag{11}$$

El algoritmo de GROMACS se muestra en la tabla 4. Las ecuaciones se resuelven simultáneamente en pequeños pasos de tiempo. El sistema se sigue durante algún tiempo, cuidando que la temperatura y la presión

permanezcan en los valores requeridos, y las coordenadas se escriban en un archivo de salida a intervalos regulares. Las coordenadas en función del tiempo representan una trayectoria del sistema. Después de los cambios iniciales, el sistema generalmente alcanzará un estado de equilibrio. Al promediar una trayectoria de equilibrio, muchas propiedades macroscópicas pueden extraerse del archivo de salida.

ALGORITMO GLOBAL DE SDM
1. Condiciones iniciales de entrada Interacción potencial *V* en función de las posiciones de los átomos. Posiciones *r* de todos los átomos en el sistema. Velocidades *v* de todos los átomos en el sistema ⇓
repita 2,3,4 para el número requerido de pasos:
2. Calcular las fuerzas La fuerza sobre cualquier átomo: $F_i = \frac{\partial V}{\partial r_i}$ se genera calculando la fuerza entre átomos no enlazados (no covalentes): $F_i = \sum_j F_{ij}$ Más las fuerzas debidas a interacciones covalentes (que pueden depender de 1,2, 3 o 4 átomos), más fuerzas de restricción y / o externas. Las energías cinética y potencial y el tensor de presión pueden ser calculados. ⇓
3. **Actualizar configuración** El movimiento de los átomos se simula resolviendo numéricamente las ecuaciones de movimiento de Newton
4. Si es necesario: resultados Escribir posiciones, velocidades, energía, temperatura, presión, etc.

Tabla 4. Algoritmo de GROMACS.

Podemos considerar que hay cuatro pasos diferentes en una simulación de SDM:

1. Configuración del sistema:
   Esta etapa incluye la mayoría de los aspectos mencionados en las secciones anteriores. Tenemos que crear una configuración razonable del sistema en el que estamos interesados, elegir un campo de fuerza válido para este sistema, seleccionar un radio de corte y el método para tratar las interacciones electrostáticas si el sistema contiene cargas parciales, elegir un algoritmo de integración (si el código de SDM que usamos lo permite) y un paso de tiempo correcto, y seleccione el conjunto de trabajo (NVE, NVT o NPT).
2. Equilibrio:
   Normalmente, la configuración inicial no será representativa de las condiciones que queremos explorar, si las posiciones iniciales se determinaran de forma aleatoria, es posible que tengamos algunos átomos demasiado cercanos o que no existan correlaciones de corto alcance, lo que no es real. Este paso generalmente requiere el uso de una simulación NPT para permitir que el sistema alcance la densidad de equilibrio correspondiente a la presión y temperatura deseadas. Es importante garantizar que hemos alcanzado un estado equilibrado, por lo que durante esta fase debemos seguir la presión, la densidad y los diferentes componentes energéticos, que en el equilibrio deben fluctuar alrededor de algún valor promedio sin mostrar ninguna desviación.

3. Producción:
   Una vez alcanzado una situación de equilibrio a la temperatura y presión deseadas, se procede a la producción de la SDM. Se debe decidir cuánto tiempo se debe ejecutar la simulación; si se estudia una biomolécula grande que puede adoptar diferentes conformaciones, el tiempo de simulación debe ser suficiente para permitir que la macromolécula explore todas las configuraciones posibles. De lo contrario, los resultados obtenidos pueden corresponder a una de las conformaciones globales de la macromolécula, pero no a la situación real.
4. Análisis:
   Finalmente, la trayectoria simulada debe ser analizada para extraer las propiedades deseadas. Se tiene acceso a las posiciones atómicas, velocidades e incluso fuerzas en función del tiempo, por lo que se puede calcular cualquier propiedad estadística mecánica que pueda expresarse en términos de esas variables.

Justificación

La proteína CEMP1 participa en la formación de los tejidos mineralizados, actuando favorablemente en la nucleación de fosfatos de calcio. Sin embargo, los mecanismos de interacción específicos entre CEMP1 y la hidroxiapatita (HA) aún son desconocidos. Actualmente, se desconocen cuáles son las regiones o dominios dentro de la secuencia de la proteína CEMP1 que actúan en la mineralización.

Existen distintas metodologías que nos permiten resolver la estructura tridimensional de las proteínas, como difracción de rayos X y resonancia magnética nuclear (NMR). Sin embargo, solo la NMR en estado sólido (SS-NMR) permite resolver la estructura tridimensional de las proteínas unidas a una superficie mineral y determinar las interacciones específicas proteína-mineral. Una de las problemáticas de SS-NMR, es que solo se pueden observar proteínas de bajo peso molecular o pequeños segmentos de proteína. Además, escanear distintas regiones de una proteína hasta completarla es una labor titánica, asumiendo que el equipo se encontrará en el país y contáramos con ilimitados recursos.

Debido a ello, una mejor opción para obtener un panorama *atómico* de las interacciones proteína-mineral es realizar estudios *in silico* como las simulaciones de dinámica molecular. Las simulaciones de dinámica molecular (SDM) permiten realizar una exploración funcional de las proteínas en un corto tiempo, con limitados recursos y de manera eficiente. La SDM permite identificar la naturaleza química de la unión CEMP1-HA, así como la identificación de dominios mayor capacidad de unión a HA. Los resultados adquieren relevancia para el diseño inteligente de péptidos con potencial terapéutico y ser comparados con pruebas experimentales. De manera importante los resultados de SMD permiten acortar las pruebas experimentales a condiciones particulares, como consecuencia se ahorran recursos, tiempo y se favorecen resultados exitosos.

Hipótesis

La Proteína CEMP1 contiene entre un 35 y 45% de aminoácidos con una estructura al azar, se encuentra relacionada a la formación del cemento radicular debido a su capacidad de unirse a sus ligando fisiológicos, como la HA. Estas interacciones requieren de aminoácidos cargados positiva y negativamente, así como también de las fosforilaciones. Las interacciones entre CEMP1 y su ligando fisiológico HA son guiadas principalmente por las fuerzas electrostáticas.

Objetivo

Identificar los motivos de unión y la estructura de la proteína CEMP1 a su ligando fisiológico (HA).

Objetivos específicos:

- Identificar los motivos en secuencia de aminoácidos en CEMP1 en función de su flexibilidad y su carga eléctrica, usando distintos algoritmos de bioinformática.
- Determinar los dominios de unión a HA en la proteína CEMP1 mediante un método de detección de péptidos orientados en forma paralela a una hojuela de HA en simulaciones de dinámica molecular.
- Identificar el papel de las fosforilaciones en CEMP1 en la asociación a su ligando y en la formación de estructura secundaria.

Material y Métodos

1. Material

Bases de datos

- Genbank: esta base de datos del Centro Nacional para la Información sobre Biotecnología (NCBI), contiene todas las secuencias de ADN disponibles públicamente, con anotaciones que incluyen la identificación de los genes, los productos de los genes (si se conocen), y las conexiones a toda clase de información sobre en gen en otras bases de datos.
- PDB (Protein Data Bank): Contiene todas las estructuras tridimensionales de proteínas y de ácidos nucleicos, resueltos por cristalografía de Rayos X y por NMR.

2. Métodos

2.1 Análisis bioinformático

- Búsqueda de secuencia de aminoácidos de CEMP1

Se utilizó la base de datos de Genbank, para la búsqueda de la secuencia de aminoácidos que componen a la proteína CEMP1.

- Búsqueda de homólogos en el genoma

Usando el programa BLAST (McGinnis and Madden, 2004) y la base de datos Genome: Homo sapiens GRCh38.p12 [GCF_000001405.38]. Se realizó una búsqueda de genes homólogos (parálogos) en el genoma humano. La búsqueda se realizó con los siguientes ajustes: Blast Genoma: Humano, con la página de búsqueda "blastn", ingresando la secuencia de nucleótidos que codifican para CEMP1 en formato FASTA, formato informático basado en texto, para representar las secuencias de ácidos nucleicos o aminoácidos; se utilizó el programa de búsqueda "Megablast" (secuencias muy similares) Tabla 2.

- Análisis de secuencia de aminoácidos

Para analizar la secuencia de aminoácidos de CEMP1 se determinó el contenido y posición de aminoácidos con carga positiva y negativa, tripletes ácidos y aminoácidos capaces de fosforilarse (NetPhos (Kreegipuu et al., 1999)), estructura secundaria (CFSSP (Chou, Peter Y. and Fasman, Gereald D., 1974), GOR IV (Kouza et al., 2017) y SOPMA (Geourjon and Deléage, 1995)) y contenido de estructura desordenada (PONDR (Ferron et al., 2006).

- Análisis de secuencia de aminoácidos con diferentes herramientas bioinformáticas

Se utilizó PONDR (Ferron et al., 2006) para predicción del contenido de estructura desordenada con la extensión VL-XT; NetPhos para la predicción de fosforilaciones; a través del servidor de ExPASy se buscaron herramientas para la predicción de estructura secundaria de las cuales se utilizaron CFSSP (Chou, Peter Y. and Fasman, Gereald D., 1974) GOR IV (Kouza et al., 2017)y SOPMA (Geourjon and Deléage, 1995); para todas estas herramientas se ingresó como "output" la secuencia de aminoácidos en formato FASTA.

2.2 Construcción de modelos

- Construcción de los modelos de los péptidos

Con base en el análisis bioinformático se procedió a la selección de los péptidos y construcción de los modelos lineales de aminoácidos mediante el programa de PyMOL (Delano, W.L., 2002) ; de manera que se colocan extremos CAP-a-a-CAP y se añade aminoácido por aminoácido.

La elaboración de los modelos fosforilados se llevó acabo con el uso del paquete de programas "YASARA" ((Krieger and Vriend, 2014) que nos permite mutar los aminoácidos de serina, treonina y tirosina por las variantes fosforiladas de fosfoserina (SEP) y fosfotreonina (TPO).

- Construcción de los modelos de HAP

El cristal de HA tiene una estructura hexagonal con un grupo espacial P63/m. Los parámetros de la celda unitaria, son a=9,424A°, b= 9,424A°, c= 6.853A°, alfa= 90°, beta= 90°, gamma=120°. Se construyó una supercelda de HA (Figura 8) de acuerdo con Mostafa et al., usando el programa Materials Studio para la construcción y optimización de la geometría de las superceldas HA, en el eje cristalográfico (100).

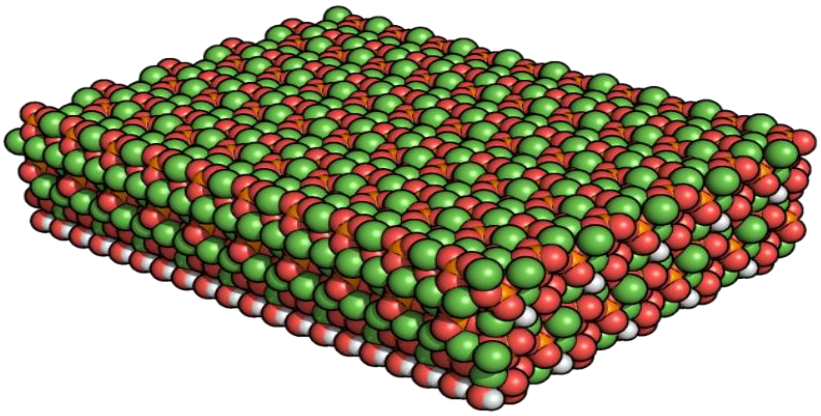


Figura 8. Esquema de supercelda de HA en eje cristalográfico (100).

2.3 Simulaciones de dinámica molecular

Se utilizó el paquete de programas de GROMACS versión 5.0.4 y el campo de fuerza de GROMOS96 43a1p complementado con los parámetros de fuerza para HA para la realización de las simulaciones de dinámica molecular. Se utilizó el modelo de agua SPC (por sus siglas en ingles Simple Point Charge) que define un modelo de tres cargas puntuales. Se utilizaron condiciones periódicas en las cajas de simulación con una geometría cubica. Para equilibrar la carga neta del sistema se agregaron iones de Na^+ o Cl^- de acuerdo a lo requerido por cada sistema.

Para neutralizar la carga del sistema, se realizaron minimizaciones de energía para una fuerza máxima <1000 kJ/mol/nm, con la finalidad de asegurar que el sistema tenga una estructura con inicio adecuada en términos de geometría y orientación del solvente; para equilibrar el solvente y los iones alrededor del péptido se llevó a cabo mediante dos fases, la primera se realizó el ensamblaje de conjunto de NVT (Número constante de partículas, Volumen y Temperatura), también conocido como "canónico" o "isotérmico-isocórico" a temperatura constante de 300K y por un periodo de tiempo de 100ps, con este paso estabilizamos la temperatura del sistema; mediante la segunda fase estabilizamos la presión (y, por lo tanto también la densidad), esto se realizó bajo un conjunto NPT (Número de partículas, Presión y la Temperatura son constantes), también se denomina conjunto "isotérmico-isobárico" a temperatura constante de 300K, presión de bar -1 y por un periodo de tiempo 100ps.

Los sistemas para las simulaciones de dinámica se construyeron de la siguiente manera:

* Péptidos sin fosforilar
- Cada péptido se colocó centrado en la celda unitaria cúbica a una distancia mínima de 1nm del borde de la caja.
- Generación de topología, posiciones de restricción.
- Solvatación con moléculas de agua utilizando el modelo SPC.
- Adición de iones necesarios para cada sistema (Na^+ o Cl^-) para una carga total cero.
- Minimización de energía para evitar choques estéricos y geometrías inapropiadas.
- Ensamblaje de NVT y NPT.
- Se liberan las posiciones de restricción y se lleva acabo SDM por un tiempo de 25ns.
- Análisis
  Se analizaron las siguientes variables en cada sistema: temperatura, presión, energía potencial, volumen, densidad fluctuaciones cuadráticas medias (RMSD), radio de giro, la predicción de estructura secundaria mediante DSSP y generación de gráficos de Ramachandran.

* Péptidos en presencia de HAP

- Cada péptido se colocó centrado y a una distancia de 3nm de la hojuela de HA (distancia C) de manera que esta distancia fuese la misma que la distancia del péptido al borde de la caja (distancia A) más la distancia de la hojuela al borde de la caja (distancia B) como se muestra en la figura 9.

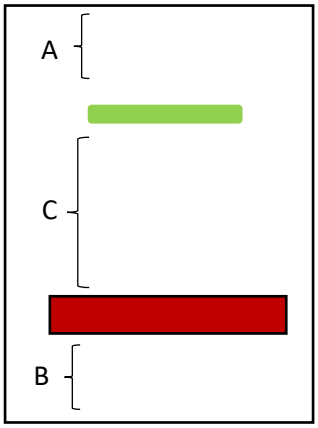


Figura 9. Construcción de los sistemas con HAP. Rectangulo rojo representa la hojuena de HA y el rectangulo verde el péptido.

- Generación de topología, posiciones de restricción (para el péptido y HA)
- Solvatación
- Adicionar iones necesarios para cada sistema (Na^+ o Cl^-) para una carga total cero.
- Minimización de energía para evitar choques estéricos y geometrías inapropiadas.
- Ensamblaje de NVT
- Se liberan las posiciones de restricción únicamente del péptido y se lleva acabo SDM por un tiempo de 25ns.

2.4 Análisis de simulaciones de dinámica molecular

Se analizaron las siguientes variables en cada sistema: temperatura, presión, energía potencial, volumen, densidad fluctuaciones cuadráticas medias (RMSD), radio de giro, formación de puentes de hidrogeno, cambio de la distancia entre el péptido y la hojuela, la predicción de estructura secundaria mediante DSSP y generación de gráficos de Ramachandran.

- Radio de Giro (Rg): Es un parámetro para estimar el grado de compactación de una macromolécula (Lobanov et al., 2008)en un solvente, midiendo la distancia que existe entre el centro de masa del péptido y el área accesible al solvente del mismo figura 10. Entre menor sea el Rg menor es la superficie que se encuentra accesible al solvente, por tanto el grado de compactación es mayor. El parámetro es relevante para determinar cambios estructurales en las macromoléculas, las proteínas globulares tienen un grado de compactación mayor en comparación a las IDPs. Los gráficos de Rg se muestran en nanómetros (nm) contra el tiempo (ps), y el valor promedio final se toma de los últimos 5 ns de la SDM

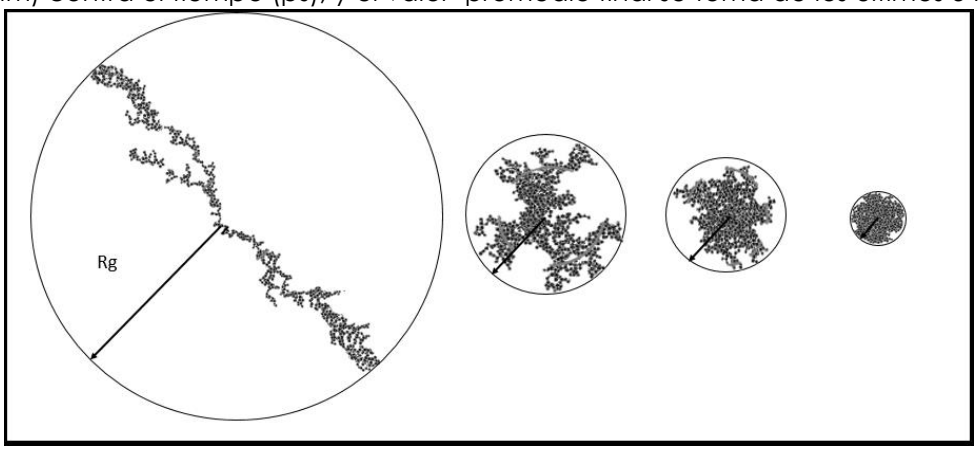


Figura 10. Radio de giro tomado de (Paszun and Dominik, 2009)

- La fluctuación cuadrática media (RMSF): Es una medida estadística para determinar la movilidad de cada uno de los aminoácidos durante la SDM, y además obtener la configuración promedio de los aminoácidos.
- DSSP: Es una herramienta para determinar estructura secundaria de cada aminoácido durante la SDM.

- Gráfico de Ramachandran

El gráfico de Ramachandran nos proporciona información sobre los ángulos de torsión de la estructura de una proteína; los ángulos de torsión de una cadena polipeptídica describen las rotaciones de la cadena principal alrededor de los enlaces entre N-C alfa (llamado Phi, φ) y C alfa-C (llamado Psi, ψ), figura 11. Estos valores nos permiten determinar posiciones estructurales de los aminoácidos y son relacionados al plegado de la proteína.

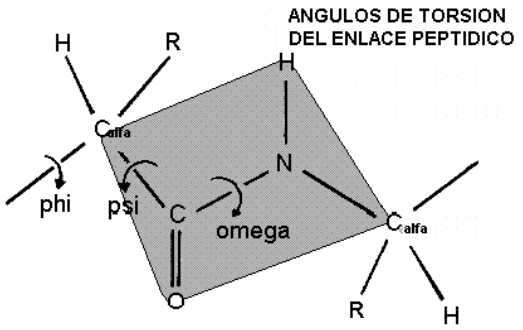


Figura 11. Ángulos de torsión del enlace peptídico. Tomado de http://www2.udec.cl/~jmartine/Capitulo3.htm (2019)

En la figura 12 se muestra el diagrama desarrollado por (Ramachandran et al., 1963)sin paréntesis; este diagrama es una forma de visualizar los ángulos diedros φ contra ψ de los residuos de aminoácidos en la estructura de la proteína. Ramachandran uso modelos de pequeños polipéptidos para variar sistemáticamente estos ángulos con el objeto de encontrar conformaciones estables; para esto se consideró a los átomos como esferas rígidas con dimensiones correspondientes a su radio de van der Waals. Así reconoció que muchas combinaciones de ángulos en una cadena polipeptídica están prohibidas debido a las colisiones estéricas entre los átomos. Su gráfico bidimensional muestra los valores permitidos y no permitidos de φ y ψ: tres cuartas partes de las combinaciones posibles están excluidas simplemente por los choques estéricos locales. La exclusión estética es el hecho de que dos átomos no pueden estar en el mismo lugar al mismo tiempo.

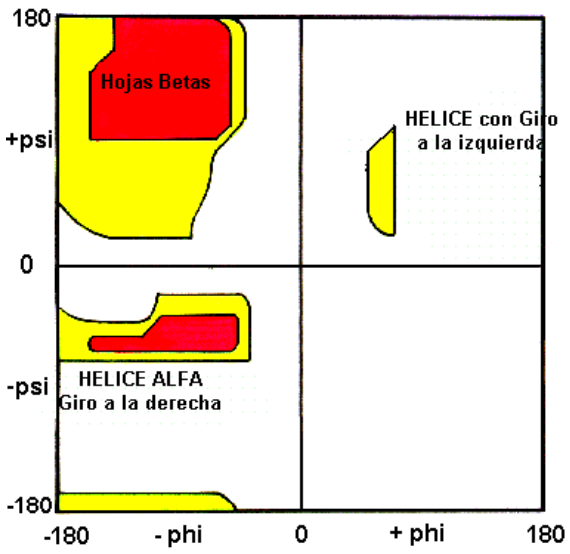


Figura 12. Gráfico de Ramachandran. Tomado de http://www2.udec.cl/~jmartine/Capitulo3.htm (2019).

En la figura 12 las zonas blancas corresponden a conformaciones donde los átomos del polipéptido están más cercanos que sus radios de van der Waals. Estas regiones son estéricamente no permitidas para todos los aminoácidos, excepto para glicina, debido a su corta cadena lateral que le permite una gran flexibilidad así

como ángulos de torsión, otro aminoácido con propiedades especiales en términos de sus ángulos de torsión es la prolina, que fija ángulos y se encuentra a menudo al final de las hélices funcionando como un disruptor. Las regiones rojas corresponden a conformaciones donde no hay impedimentos estéricos, es decir estas son zonas permitidas llamadas conformaciones a-hélices y lámina-β. Las zonas amarillas muestran las regiones permitidas sí en los cálculos se usan radios van der Waals ligeramente más pequeños, es decir se permite que los átomos puedan estar un poco más cercanos; esto genera una nueva región que corresponde a una hélice levógira (giro a la izquierda).

La obtención de gráficos de Ramachandran se realizó para todos los péptidos así como la cuantificación de los ángulos presentes correspondientes a estructura α hélice dextrógira (φ -60°,ψ-40°), α hélice levógira (φ 60°, ψ 40°) y lámina β (φ-140°, ψ135°); para todos los ángulos se consideró un rango de ⁺∕₋ 10° (Hollingsworth and Karplus, 2010).

- Puentes de hidrógeno

La estructura secundaria en una cadena polipeptídica se adopta principalmente gracias a la formación de puentes de hidrógeno. Los puentes de hidrógeno se establecen entre los grupos –CO- (aceptor de H) y –NH- (donador de H) del enlace peptídico; de esta forma la cadena polipeptídica es capaz de adoptar una conformación de menor energía y más estable.

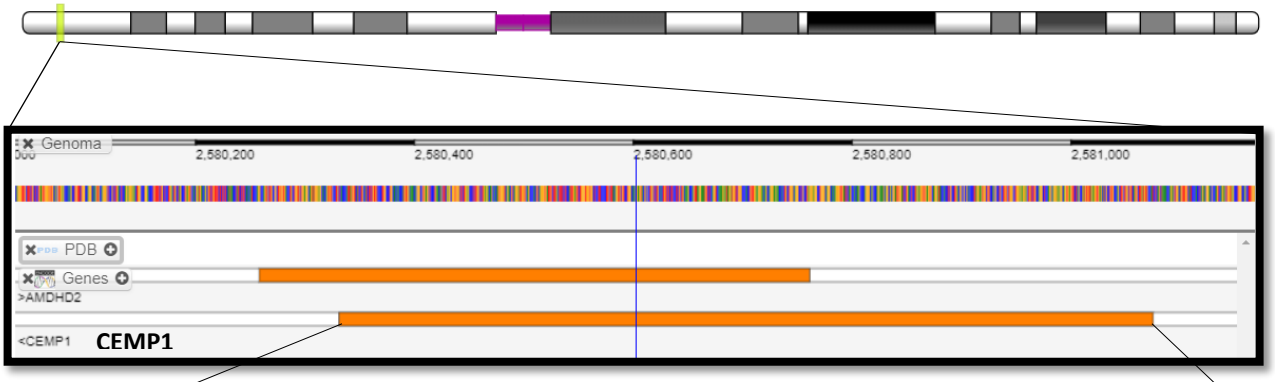
- Distancia entre HA y péptido

En los sistemas de mide la distancia que existe entre el centro de masa del péptido y el centro de masa de la hojuela de HA, durante todo el tiempo de dinámica. Para este parámetro se calculó la diferencia entre la distancia inicial y final entre el péptido-HA de esta manera se puede determinar fácilmente si el péptido se acerca o aleja a la hojuela de HA.

Resultados

1. Análisis bioinformático de CEMP1

El gen con número 752014 en la base de datos Genbank corresponde a la Proteína de Cemento 1 [*Homo sapiens* (humano)] y se localiza en el brazo corto del cromosoma 16 en el locus 13.3, no posee péptido señal y la secuencia de nucleótidos que la codifican se muestra en la Figura 13.



>NC_000016.10:c2531408-2530035 Homo sapiens chromosome 16, GRCh38.p12 Primary Assembly
TTACTTTTGTTATGTATTTTACCACAATTTTTAAAAAGCAAACCAAAACCAACCAAGAGTGCTTCCCCCA
CACCTCAAAATCATCCTGCAGCAGCTCCCTGGCCCAGCTCTCTCTCACCCTGACCCTGGGCCCCTCTCCC
ACCACCCAGGGCTAGCCCTGTGGACCAACCATCTCTGCCAGCCCCTCCCCGACCCTCCAGCCAGGGAGGT
GGGGCGCTGGCCGGTGAATGGGGCAGGCCAGGCCCAAAGGCTGGCCCAGGGCTCACCAGCTCTGGACTGG
GCGTCCCGTCTGAGGTGGGGATGACCAACATGCCAGCTCTGGGTTTTAGCTGAGGATGGGCACATCAAGC
ACTGACAGCCAGCAAGCTGGGCACAGGCGATGCTCAACCTCTAACACATCGGCTGAGAACCTCACCTGCC
TCTCCCTGCCTGGGAGCCCTGGCAAGACAGCTCCTCTCCCAGGCCCAGCACAGGCAGGCGCTGGACAGCC
ACTTCCTAAGGGGTGCGCGGCCGTCAAGGCAGAGGTGGGTATCCCTGCACCTCACACCTCTCAGGAGGTC
AGGATCCACATCCGTCGCCTGCTTTCCTGGGCTGCCCCTGGGGCTTGTGGGCTGAGATCAACCCCTTGTG
CCCTGCCCCAGGCCCTCCCTCAGGCGAGGCCCTGCCCAGGCAGGTGGTTCTTCCCTGGCTGTTCCCTTCC
AACAGGTGGGGCCCAGACCATCCTATCTCTTTGGACCTGGAGACATTTCCTTAACTGGGCTTTGCAGCAG
AGGGAAGAGAACAGTGGCAGAGCCAGGCGCGTACCTCCTGTGCCCAGGACAGCACCTGTAAGCAAGGGGG
AAGGGAGCCACCCACCCCAGAACTCAAATGGAGAGAAAGTCAAGACAATCACCCCTGACGTGGGGCTGCA
CCAATCCCTGACGTCTGACCCCACTGTGGCTGTCCTCAGAGCCAAGAGGGCTCCAGAAGCCCACCCGCCT
CGAAGCTGTTCAGGGAGCCTCACGGCCCGTGTTTGCCACATGGGTGTGTGCCAGGGCCAAGGGGACACAG
AAGATGGCAGGATGACACTAATGGGGTGAGCAGGGCACACTGGTGCCTCCCATACTGCCCTCCAGCACCA
GGCAGGGTCACCACAGCCGGCAAGCACTGGGCCTGGGGATGTGTGAAGAGATAGGGTTATTGAGAAGAGA
AAACAGAACGCTGGTCATCCGTGGCACTCAGGGCTAGGCAGTCCAGGGAGCAGAAAACAGGGGTCCCGGG
CCCAGTCTCCCTCCTGGAGGTCAGAGCGGAGCCCCTGGCACTGTGCCCTGGCCCAGAACCAGGCTGTGGG
CTCCACGCTGCACACCCCAACTCAGGAGCAGAAAGCAAACCCCT

Figura 13. Resultado de busqueda de CEMP1 en la base de datos de Genbanck.

De acuerdo con nuestra búsqueda la proteína de CEMP1 no tiene homólogos en el genoma y su mejor correlación es con ella misma, resultados del programa BLAST se muestra en la Figura 14 y 15.

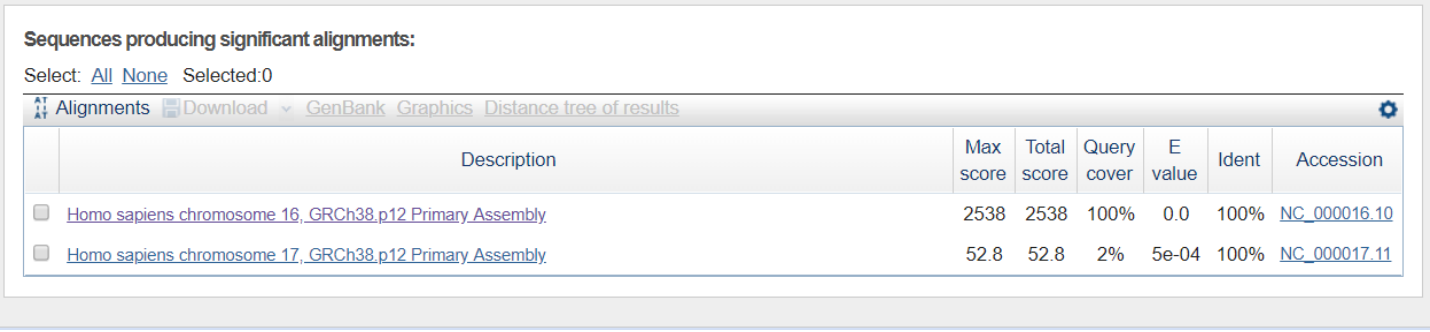


Figura 14. Secuencias que producen alineaciones significativas en el Blast genómico de CEMP1.

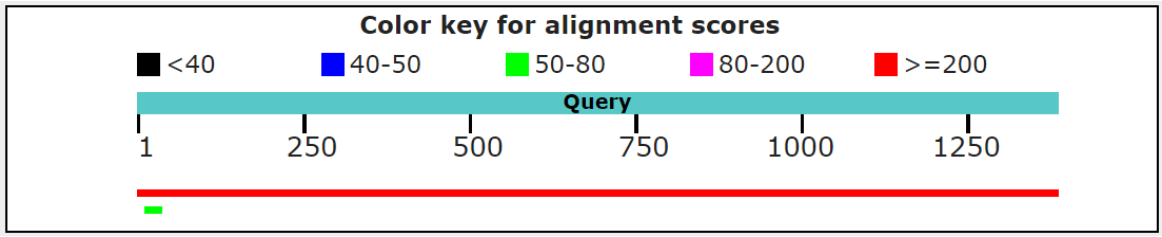


Figura 15. Resultado gráfico de Blast genómico de CEMP1.

La secuencia de aminoácidos de CEMP1 cuadro 1, fue obtenida a partir de la búsqueda en la base de datos Genbank. La proteína está compuesta por 247 aminoácidos.

MGTSSTDSQQAGHRRCSTSNTSAENLTCLSLPGSPGKTAPLPGPAQAGAGQPLPKGCAAVKAEVGIPAPH
TSQEVRIHIRRLLSWAAPGACGLRSTPCALPQALPQARPCPGRWFFPGCSLPTGGAQTILSLWTWRHFLN
WALQQREENSGRARRVPPVPRTAPVSKGEGSHPPQNSNGEKVKTITPDVGLHQSLTSDPTVAVLRAKRAP
EAHPPRSCSGSLTARVCHMGVCQGQGDTEDGRMTLMG

Cuadro 1. Secuencia de aminoácidos de CEMP1

* Análisis de la secuencia de aminoácidos

CEMP1 cuenta con un gran número de aminoácidos cargados positiva (histidina, arginina, lisina) y negativamente (ácido aspártico y ácido glutámico), tripletes acídicos, así como aminoácidos susceptibles a fosforilación como serinas, treoninas y no posee tirosinas. Tablas 5-7.

Aminoácido	*Nombre*	*N° de a.a. en secuencia*	*Posición en secuencia*
Positivos	Histidina [H]	8	13,70,78,137,172,202,213,228
	Arginina [R]	19	14,15,76,80,81,94,108,113,136,146,152,154,155,161,205,208, 216,225,242
	Lisina [K]	7	37,55,61,167,181,183,197
Negativos	Ácido Aspártico [D]	5	7,188,198,237,240
	Ácido glutámico [E]	9	24,63,74,147,148,169,180,211,239

Tabla 5. Aminoácidos positivos y negativos de CEMP1.

TRIPLETES ÁCIDOS		
TRIPLETE	N°	POSICIÓN EN SECUENCIA
TDS	1	6-8
TSD	1	196-198
DTED	1	237-240

Tabla 6. Tripletes acídicos presentes en CEMP1.

Aminoácido	**N° de a.a. en secuencia**	**Posición en secuencia**
Serina [S]	22	4,5,8,17,19,22,30,34,72,84,95,120,131,150,166,171,177,194,197,217,219,221
Treonina [T]	19	3,6,18,21,27,38,71,96,123,127,134,162,184,186,196,200,223,238,244
Tirosina [Y]	0	

Tabla 7. Aminoácidos que pueden fosforilarse presentes en secuencia de CEMP1.

* Predicción de contenido de estructura secundaria

En la Figura 16 se observa la predicción de estructura secundaria de los diferentes predictores utilizados.

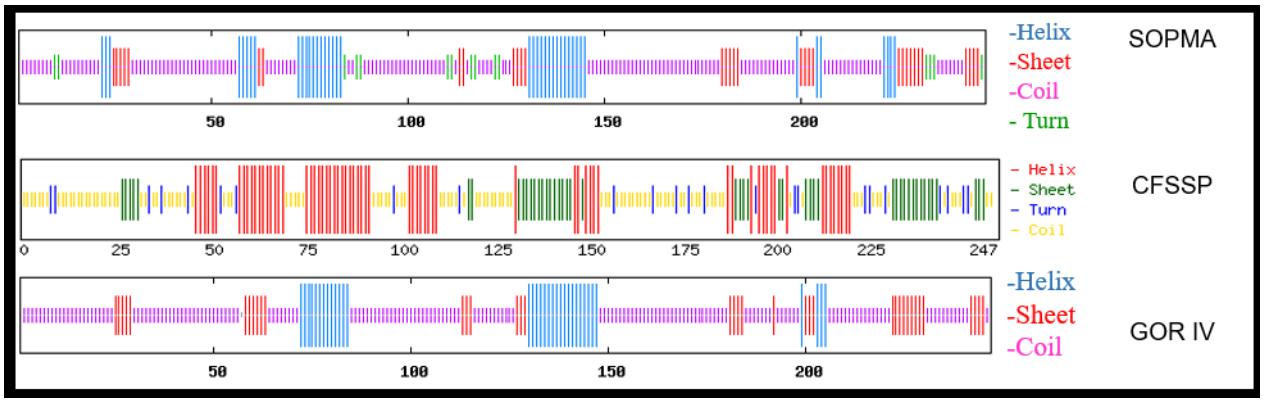


Figura 16. Resultados gráficos de los diferentes predictores de estructura secundaria.

De acuerdo con los resultados la proteína CEMP1 presenta un promedio de 35% de estructura secundaria, principalmente estructura α-hélice; además de un porcentaje promedio de estructura desordenada de 64.6%. Los resultados se muestran en la tabla 8.

Predictor	*α-Hélice (n° - %)*	*Lámina-β (n° - %)*	*% total Ordenado*	*Desordenado (n° - %)*
SOPMA	42 – 17%	35 – 14.17%	31.17%	170 – 68.82%
CFSSP	67 – 27.12%	46 – 18.62%	45.74%	134 – 54.24%
GOR IV	48 – 19.43%	24 – 9.71%	29.14%	175 – 70.85%

Tabla 8. Resultados de predicción de estructura secundaria. Para cada uno de los tipos de estructura se presenta del lado izquierdo el n° de residuos que corresponden a esa estructura, seguido del porcentaje correspondiente.

* Predicción de estructura ordenada y desordenada

Los resultados de PONDR se muestran en la Figura 17 y resumidos en la tabla 9. CEMP1 presenta un porcentaje total de estructura desordenada del 47.77%. De los 247 aminoácidos presentes en la secuencia de CEMP1, 118 se encuentran en un estado desordenado y distribuidos en 8 diferentes regiones en la proteína, siendo la región desordenada con mayor longitud de 34 aminoácidos.

Número de residuos: 247	Región desordenada más larga: 34
Número de Regiones Desordenadas: 8	Porcentaje total desordenado: 47.77
Número de residuos desordenados: 118	Puntuación de Predicción Promedio: 0.4882

Tabla 9. Resultados del predictor PONDR.

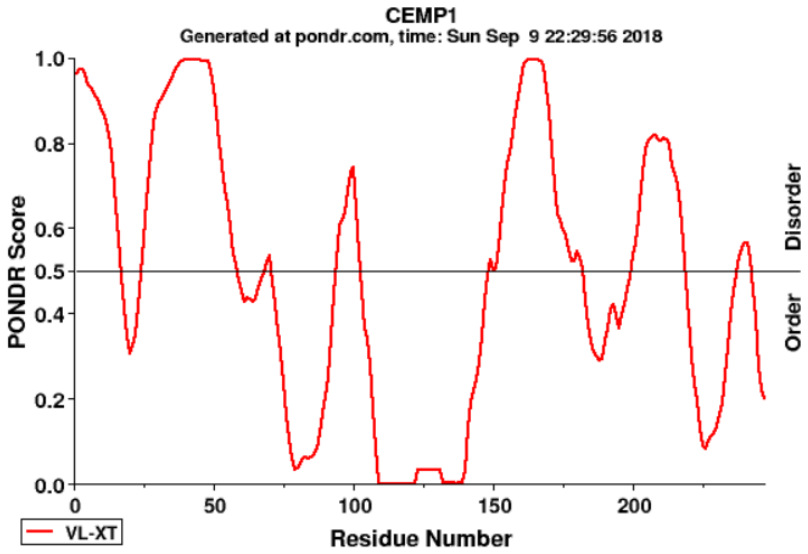


Figura 17. Resultado gráfico del servidor de PONDR de la secuencia de aminoácidos de CEMP1.

* Predicción de posibles sitios de fosforilación en CEMP1

De las 22 serinas presentes en la secuencia de CEMP1, solo 19 presentaron una puntuación mayor o igual de 0.5, el umbral para considerar una posible fosforilación. De las 19 treoninas presentes en la secuencia de CEMP1, solo 12 son consideradas para ser fosforiladas. Los resultados se presentan en la tabla 10 y Figura 18.

Aminoácido	N° de aminoácidos en secuencia	Posibles Fosforilaciones
Serina	22	19 [5,8,17,18,19,22,30,34,72,84,120,131,150,166,194,197,217,219,221]
Treonina	19	12 [6,21,71,96,123,128,134,162,186,196,223,244]
Tirosina	0	0

Tabla 10. Posibles fosforilaciones presentes en la secuencia de aminoácidos de CEMP1.

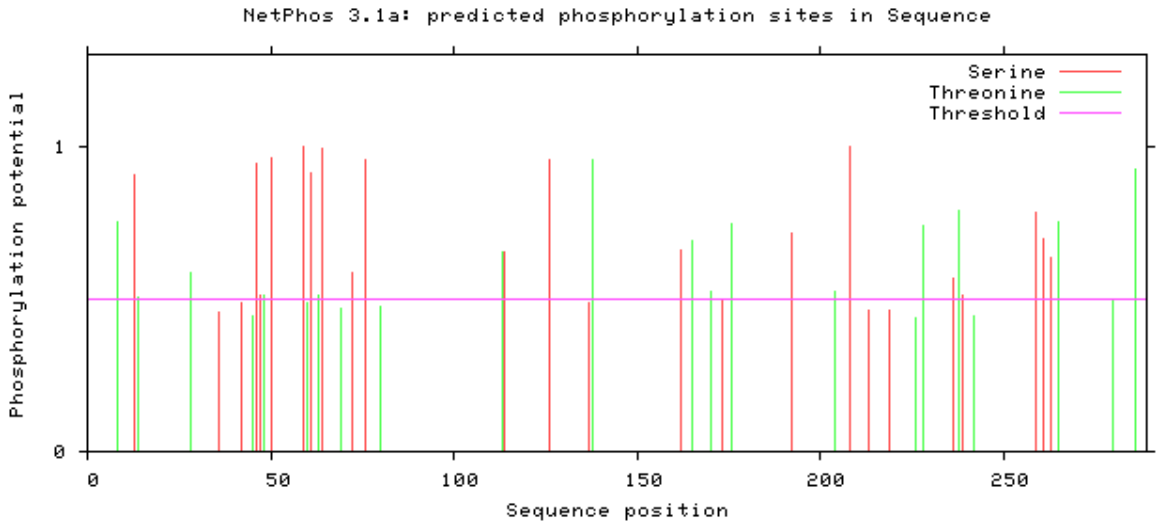


Figura 18. Resultado gráfico de NetPhos de la predicción de CEMP1.

De los 247 aminoácidos que componen a la proteína del cemento 1 (CEMP1), nos limitamos a utilizar únicamente los primeros 90 aminoácidos que componen al mismo cuadro 2, debido a que el resto de la secuencia se encuentra bajo proceso de patente.

MGTSSTDSQQAGHRRCSTSNTSAENLTCLSLPGSPGKTAPLPGPAQAGAGQPLPKGCA
AVKAEVGIPAPHTSQEVRIHIRRLLSWAAPGACGLRSTPCALPQALPQARPCPGRWFFP
GCSLPTGGAQTILSLWTWRHFLNWALQQREENSGRARRVPPVPRTAPVSKGEGSHPPQ
NSNGEKVKTITPDVGLHQSLTSDPTVAVLRAKRAPEAHPPRSCSGSLTARVCHMGVCQG
QGDTEDGRMTLMG (247 aa)

Cuadro 2. Secuencia de aminoácidos de CEMP1 (En azul la parte de la proteína a utilizar).

Dentro de región en secuencia de 1 a 90, Cuadro 3, CEMP1 cuenta con aminoácidos con cargas tanto positivas (H, R y K) como negativas (D y E), el número y posición de estos residuos se muestran en la Tabla 11. Los tripletes acídicos presentes se muestran en la Tabla 12 y los aminoácidos susceptibles a fosforilación como S y T presentes se muestran en la Figura 19 y resumidos en la tabla 13.

MGTSSTDSQQAGHRRCSTSNTSAENLTCLSLPGSPGKTAPLPGPAQAGAGQP
LPKGCAAVKAEVGIPAPHTSQEVRIHIRRLLSWAAPGA (90 a.a.)

Cuadro 3. Primeros 90 aa de CEMP1. Se muestran en verde los aa cargados positivamente; en rojo los aa cargados negativamente; en amarillo los aa suceptibles a fosforilación; y subrayado los tripletes ácidicos.

Aminoácido	*Nombre*	*N° de a.a*	*Posición en secuencia*
Positivos	Histidina (H)	3	13,70,78
	Arginina (R)	5	14,15,76,80,81
	Lisina (K)	3	37,55,61
Negativos	Ácido aspártico (D)	1	7
	Ácido glutámico (E)	3	24,63,74

Tabla 11. Aminoácidos con carga positiva y negativa en a.a 1-90 de CEMP1.

TRIPLETES ÁCIDOS		
TRIPLETE	N°	POSICIÓN EN SECUENCIA
TDS	1	6-8

Tabla 12. Tripletes acídicos presentes en CEMP1.

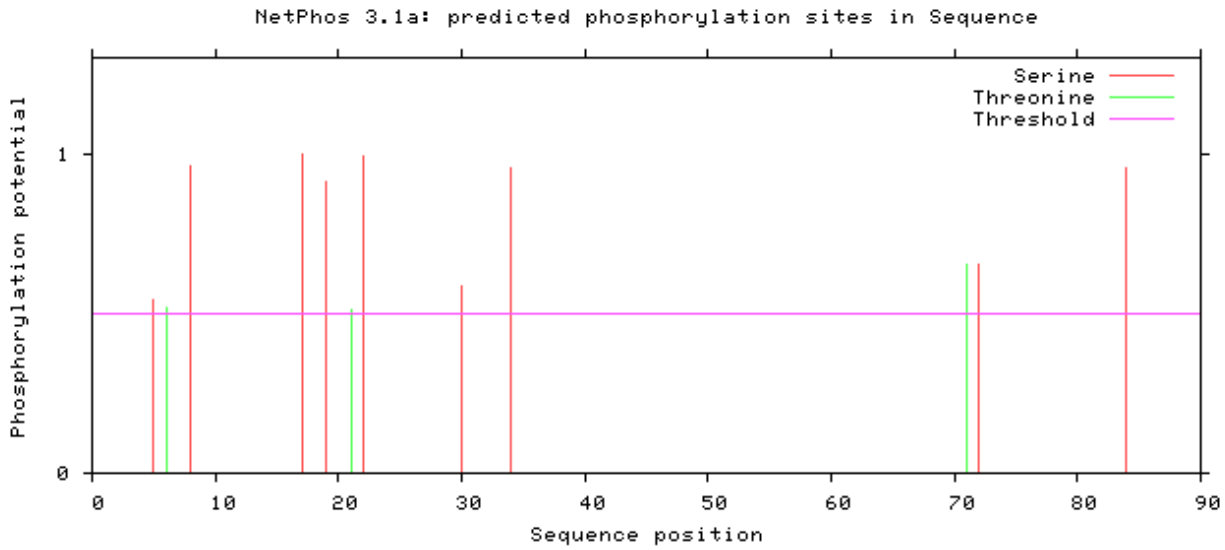


Figura 19. Resultado gráfico de predicción de fosforilaciones de NetPhos para los primeros 90 a.a de CEMP1.

Aminoácido	Susceptibles a fosforilación	N° total de sitios	Posibles fosforilaciones	N° total de sitios
S	4, 5,8,17,19,22,30,34,72,84	10	5,8,17,19,22,30,34,72,84	9
T	3,6,18,21,27,38,71	7	6,21,71	3

Tabla 13. Posibles sitios de fosforilación en CEMP1 del a.a 1-90.

De acuerdo con los resultados obtenidos de la predicción de fosforilaciones de aminoácidos de CEMP1 de las 10 serinas presentes en esta región, 9 presentaron una puntuación mayor o igual de 0.5, mientras para la treonina de las 7 presentes en la región, solo 3 presentaron una puntuación mayor, los resultados se resumen en la tabla 13.

En cuanto a estructura desordenada de los 90 residuos, los resultados de PONDR se muestran en la Figura 20. Presenta 56 residuos en estado desordenado, distribuidos en 5 regiones y siendo la más larga de 34 aminoácidos, para un porcentaje total de 62.22% de estructura desordenada, estos resultados se presentan en la tabla 14.

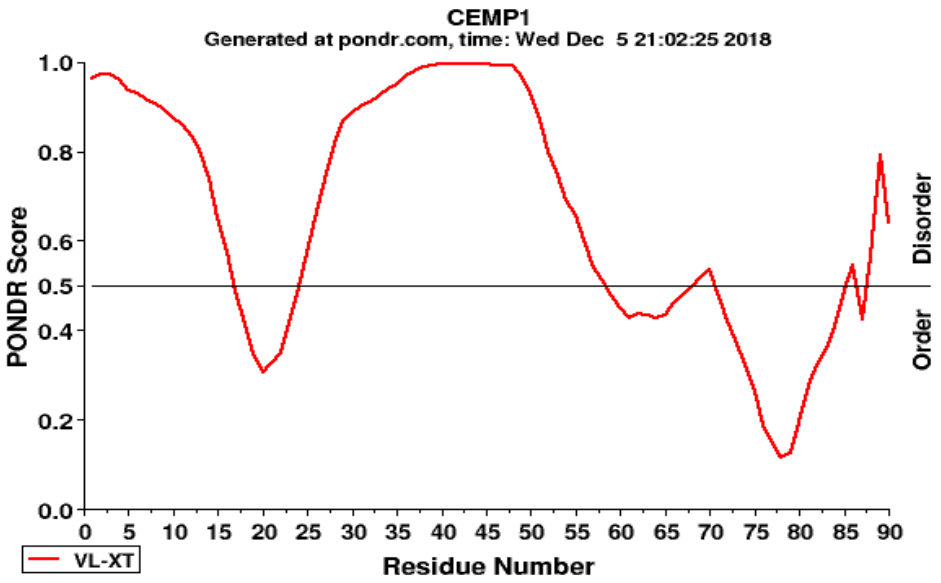


Figura 20. Resultado gráfico de PONDR de predicción de estructura desordenada en los primeros 90 aa de CEMP1.

N° de residuos: 90	*Región desordenada más larga: 34*
Regiones desordenadas: 5 (1-16, 25-58, 68-69, 86, 88-90)	Porcentaje total desordenado: 62.22
N° de residuos desordenados: 56	Score 0.5

Tabla 14 Resultados del servidor de PONDR de CEMP1 del a.a 1-90.

Esta región de la proteína CEMP1 presenta un promedio de 33.7% de estructura secundaria, principalmente estructura α-hélice; y un porcentaje de estructura desordenada de 66.29%. Los resultados se muestran en la tabla 15.

Predictor	*α-Hélice (n° - %)*	*Lámina-β (n° - %)*	*Ordenado % total*	*Desordenado (n° - %)*
SOPMA	20 – 22.22%	7 – 7.77%	30%	63 – 70%
CFSSP	35 – 38.88%	5 – 5.55%	44.44%	50 – 55.55%
GOR IV	13 – 14.44%	11 – 12.22%	26.66%	66 – 73.33%

Tabla 15. Resultados de la predicción de estructura secundaria para los primeros 90 a.a. de CEMP1.

A partir de estos datos se seleccionaron 5 diferentes péptidos que se muestran en la tabla 16, donde se observa la composición y el número total de aminoácidos de cada péptido; puesto que es de nuestro interés determinar el papel que desempeñan las fosforilaciones en la asociación al ligando HA y en la adquisición de estructura secundaria se construyeron dos versiones para cada péptido con y sin las fosforilaciones presentadas anterior mente, los residuos de aminoácidos con fosforilación se muestran en la tabla 16 en color rojo.

Péptido	Secuencia de aminoácidos	N° de aminoácidos
P1	MGTSSTDSQQAGHRRCSTSN	20
P2	TSAENLTCLSLPGSPGKTAP	20
P3	MGTSSTDSQQAGHRRCSTSNTSAENLTCLSLPGSPGKTAP	40
P4	QPLPKGCAAVKAEVGIPAPH	20
P5	TSQEVRIHIRRLLSWAAPGA	20

Tabla 16. Péptidos seleccionados.

La construcción de los modelos se realizó con el programa de PyMOL; bloqueando los extremos amino y carboxilo terminal con un grupo acetil y un grupo amino para evitar la interacción de las cargas de los grupos carboxilo y amino terminal, en la figura 21 se muestra un ejemplo del péptido 1. Las fosforilaciones se añadieron a los modelos con el paquete de programas de YASARA; el cual nos permite mutar a los aminoácidos S y T de la secuencia por sus versiones fosforiladas; se construyó el modelo completo de CEMP1 sin y con las posibles fosforilaciones indicadas por el predictor de NetPhos y se cuerdo con estas modificaciones la carga total de CEMP1 y de nuestros péptidos se muestra en la tabla 17.

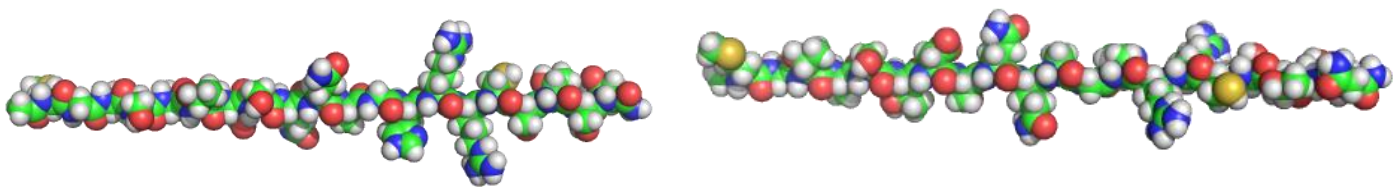


Figura 21. Modelo del péptido n°1. La imagen de lado izquierdo es el péptido sin fosforilar, lado derecho el modelo del péptido fosforilado.

Carga total (q)	
Proteína	
CEMP1 sin fosforilar	**+13**
CEMP1 fosforilada	**-47**
Péptidos sin fosforilar	
P1	+2
P2	-1
P3	+1
P4	0
P5	+3
Péptidos fosforilados	
P1F	-8
P2F	-9
P3F	-17
P4F	-----
P5F	-3

Tabla 17. Carga total de CEMP1 y de los péptidos seleccionados. P4 no presenta una versión fosforilada.

Para cada uno de los péptidos se construyeron sistemas en presencia de HA, por ende se generaron cuatro sistemas (con excepción del P4); **la nomenclatura que se muestra en la tabla 18 se utilizara en todos los resultados.**

NOMENCLATURA		
PÉPTIDO	**ACRONIMO**	**SIGNIFICADO**
1	P1	Péptido 1 sin fosforilar
	P1F	Péptido 1 fosforilado
	P1HA	Péptido 1 en presencia de HA
	P1FHA	Péptido 1 fosforilado en presencia de HA
2	P2	Péptido 2 sin fosforilar
	P2F	Péptido 2 fosforilado
	P2HA	Péptido 2 en presencia de HA
	P2FHA	Péptido 2 fosforilado en presencia de HA
3	P3	Péptido 3 sin fosforilar
	P3F	Péptido 3 fosforilado
	P3HA	Péptido 3 en presencia de HA
	P3FHA	Péptido 3 fosforilado en presencia de HA
4	P4	Péptido 4 sin fosforilar
	P4HA	Péptido 4 en presencia de HA
5	P5	Péptido 5 sin fosforilar
	P5F	Péptido 5 fosforilado
	P5HA	Péptido 5 en presencia de HA
	P5FHA	Péptido 5 fosforilado en presencia de HA

Tabla 18. Nomenclatura de los péptidos.

2. Análisis de simulaciones de dinámica molecular.

Para verificar que las simulaciones se realizaron de manera correcta se revisó el tiempo de la SDM (25ns) y la convergencia de energía en los parámetros termodinámicos de temperatura, presión, energía potencial y volumen; para cada una de las SDM. En la figura 22 se muestran una representación gráfica para la SDM del péptido 1 sin la hojuela de HA.

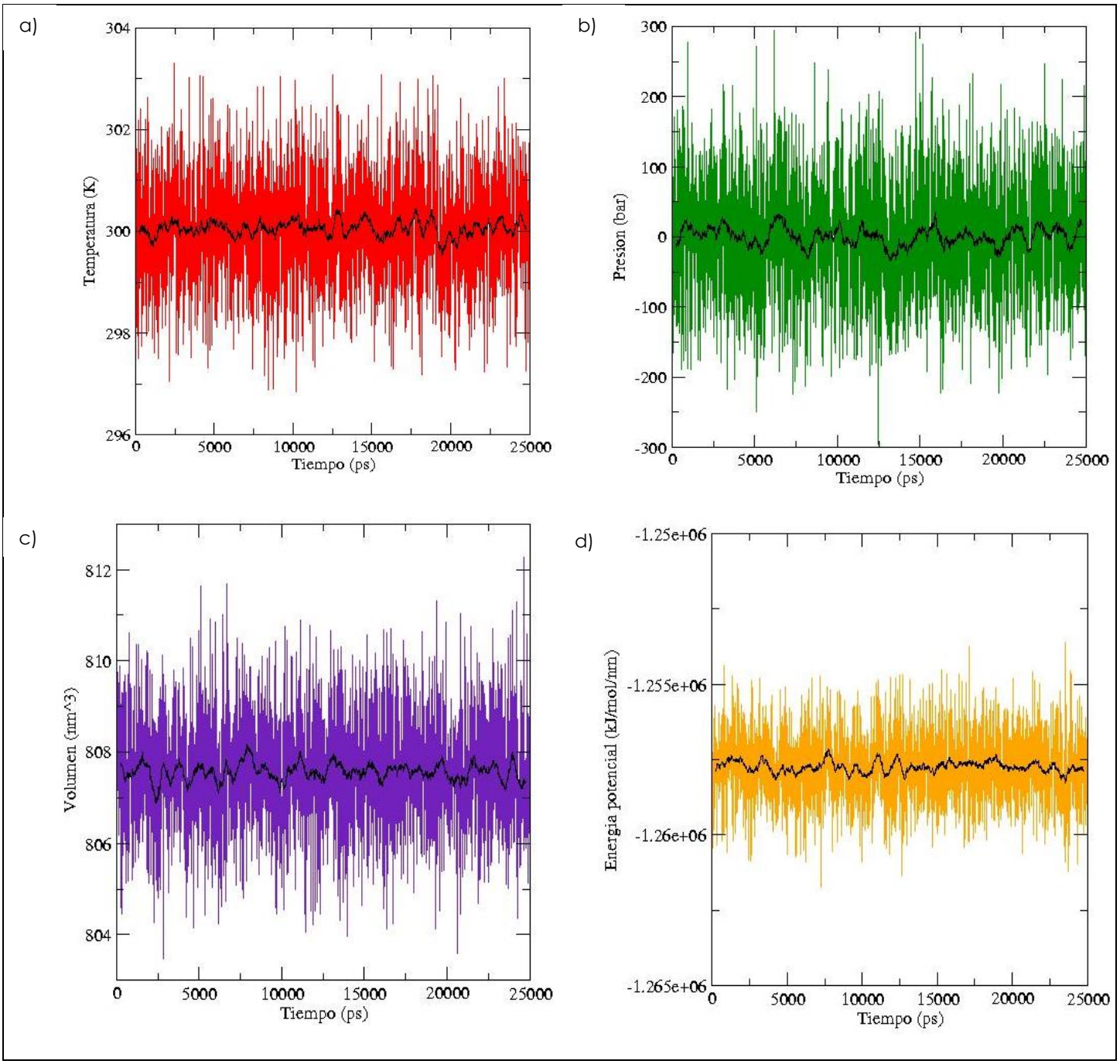


Figura 22. Gráficos de convergencia de los parámetros termodinámicos SDM péptido 1 sin HA. a) Temperatura; b) Presión; c) Volumen d) Energía potencial. Para todos los casos la línea negra indica el promedio del parámetro correspondiente en cada gráfico.

Radio de giro

- Como se muestra en la tabla 18 se construyeron cuatro sistemas para el Péptido 1. Los cambios del Rg entre los diferentes sistemas se muestran en la figura 23 y oscilan entre 0.1nm, siendo el estado más compacto P1 y el menos compacto el P1HA.

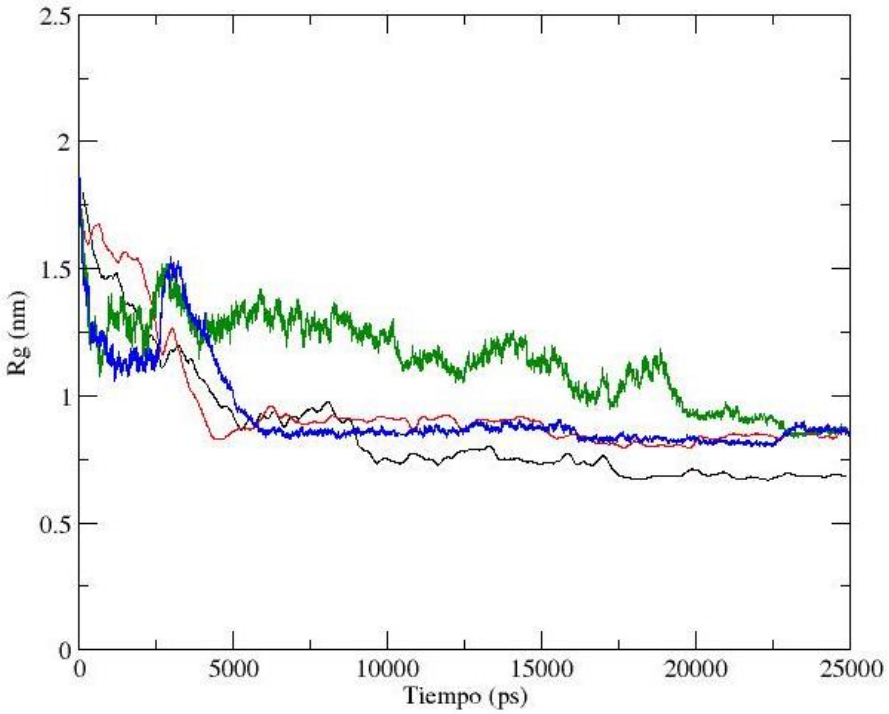


Figura 23. Rg del péptido1: P1 línea negra, P1F línea roja, P1HA línea verde y P1FHA línea azul.

- Como se muestra en la tabla 18 se construyeron cuatro sistemas para el Péptido 2. El mayor Rg es de 1.07nm y se presenta en el P2FHA, el menor Rg 0.048nm (mayor grado de compactación) se presenta en el P2HA (figura 24).

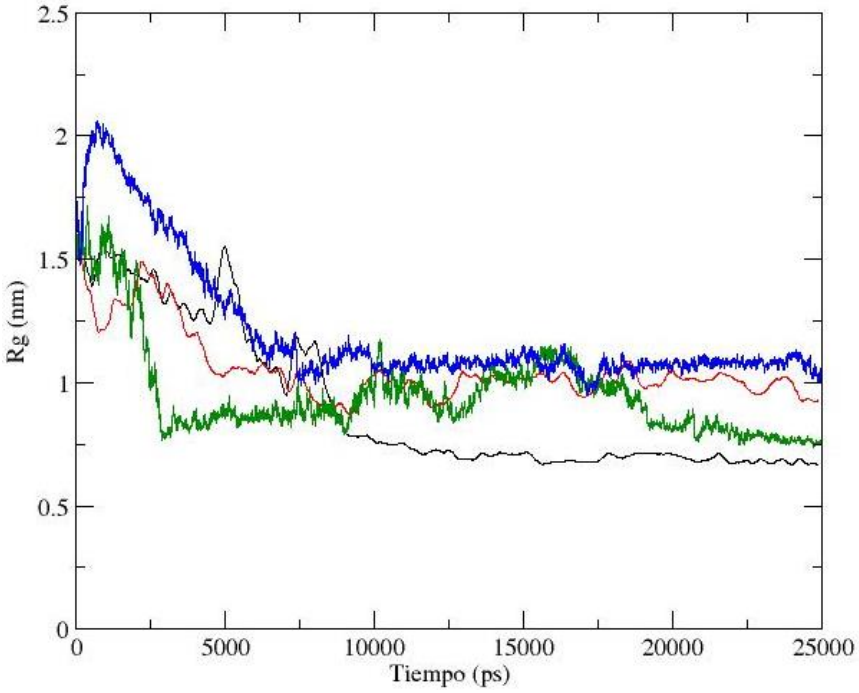


Figura 24. Rg del péptido2: P2 línea negra, P2F línea roja, P2HA línea verde y P2FHA línea azul.

- Como se muestra en la tabla 18 se construyeron cuatro sistemas para el Péptido 3. El estado más compacto (menor Rg 1nm) se presenta en los péptidos P3 Y P3F; dentro de los sistemas en presencia de HA el P3FHA muestra el menor Rg 1.60nm (figura 25).

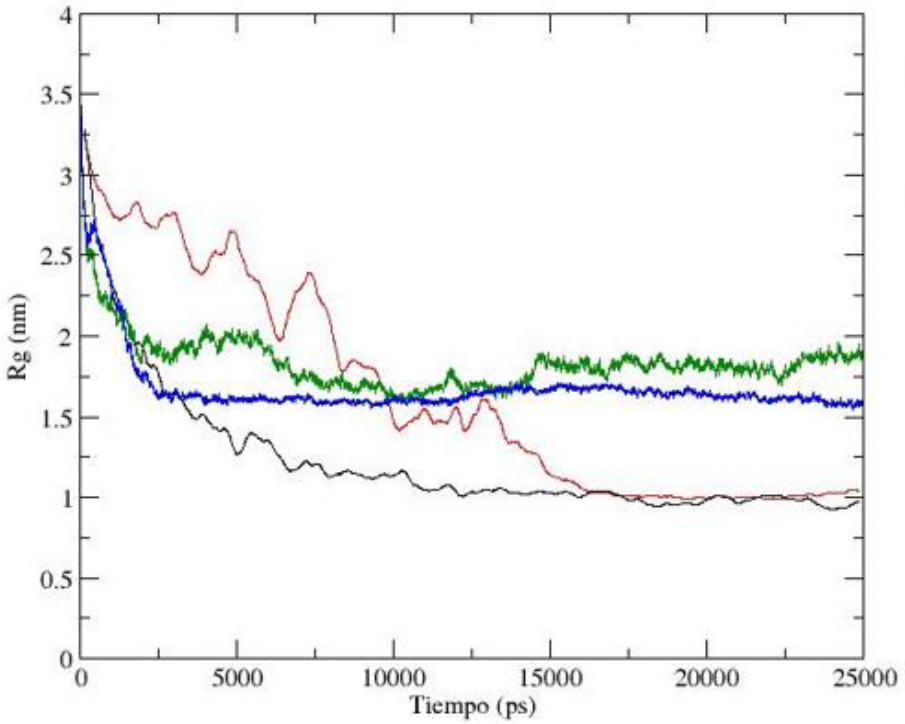


Figura 25. Rg del péptido3: P3 línea negra, P3F línea roja, P3HA línea verde y P3FHA línea azul.

- Como se muestra en la tabla 18 se construyeron dos sistemas para el Péptido 4. El cambio en el Rg entre los sistemas es de 0.7 nm, presentándose el menor Rg 0.081nm en el P4HA (figura 26).

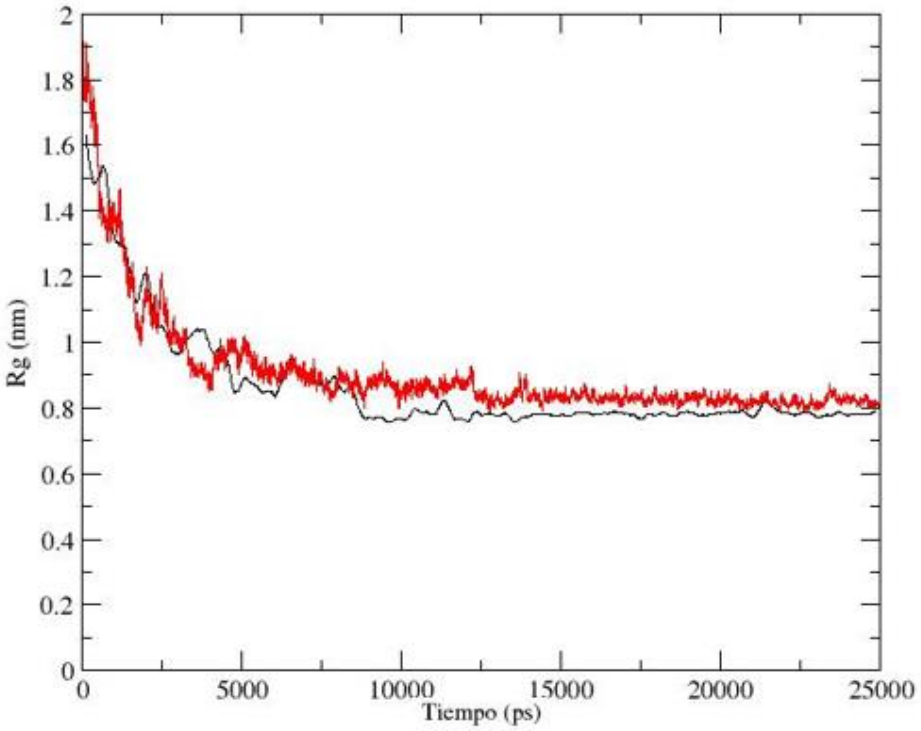


Figura 26. Rg del péptido 4: P4 línea negra y P4HA línea roja.

- Como se muestra en la tabal 18 se construyero cuatro sistemas para el Péptido 5. El Rg de los sistemas oscila entre 0.8 y 0.9 nm, presentándose menor Rg en P5 con 0.8nm y el mayor en el P5HA con 0.95 nm (figura 27).

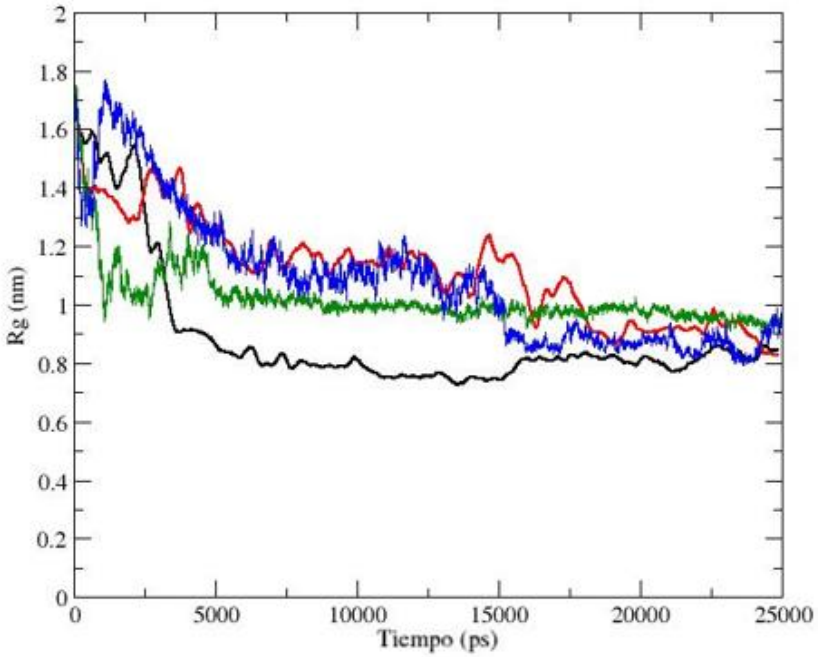


Figura 27. Rg del péptido 5: P5 línea negra, P5F línea roja, P5HA línea verde y P5FHA línea azul.

Todos los resultados del Rg asociados a la carga neta de cada péptido se muestran en la tabla 19.

Péptido	Carga total (q)	Rg (nm)
P1	+2	0.7
P1F	-8	0.8
P1HA	+2	0.863
P1FHA	-8	0.838
P2	-1	0.7
P2F	-9	1
P2HA	-1	0.048
P2FHA	-9	1.07
P3	+1	1
P3F	-17	1
P3HA	+1	1.82
P3FHA	-17	1.60
P4	0	0.8
P4HA	0	0.081
P5	+3	0.8
P5F	-3	0.9
P5HA	+3	0.957
P5FHA	-3	0.86

Tabla 19. Radio de giro

Distancia entre HA y el péptido

En los sistemas de mide la distancia que existe entre el centro de masa del péptido y el centro de masa de la hojuela de HA, durante todo el tiempo de dinámica (figura 28). Para este parámetro se calculó la diferencia entre la distancia inicial y final entre el péptido-HA de esta manera se puede determinar fácilmente si el péptido se acerca o aleja a la hojuela de HA; los resultados se muestran en la tabla 20.

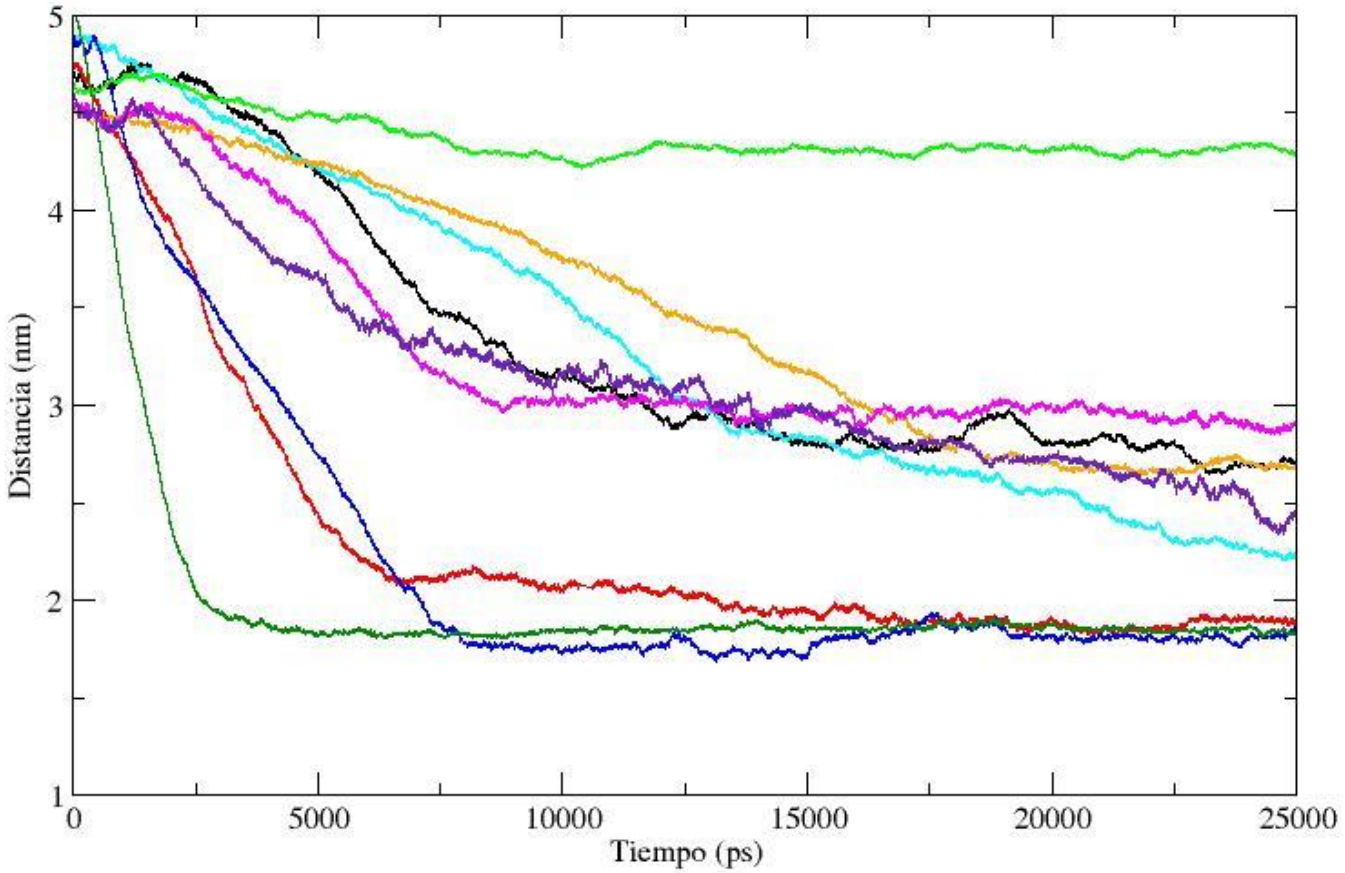


Figura 28. Distancia entre los péptidos y HA. Las lineas corresponden a: P1HA negra, P1FHA roja, P2HA azul claro, P2FHA azul marino, P3HA verde claro, P3FHA verde militar,, P4HA amarillo, P5HA rosa y P5FHA morado.

Péptido	Distancia inicial (nm)	Distancia final (nm)	Diferencia (nm)
P1HA	4.716	2.690	2.026
P1FHA	4.723	1.886	2.837
P2HA	4.883	2.244	2.639
P2FHA	4.877	1.816	3.061
P3HA	4.667	4.290	0.377
P3FHA	5.010	1.826	3.184
P4HA	4.526	2.677	1.849
P5HA	4.537	2.899	1.638
P5FHA	4.597	2.466	2.131

Tabla 20. Distancia entre HA y péptido.

RMSF

- Péptido 1

Los aminoácidos presentes con carga como el ácido aspártico (residuo 8), histidina (residuo 14), arginina (residuo 15 y 16), y aminoácidos capaces de fosforilarse como la serina (residuo 6, 9,18 y 20) y treonina (residuo 7) presentan variación en los resultados de RMSF (figura 29). Estos aminoácidos participan de manera importante en la formación de puentes de hidrogeno y observamos un incremento en cuanto a la formación de estos en la versión fosforilada donde SEP participa entre un 80-100% de estos enlaces (de acuerdo a cada sistema) ver tabla 22.

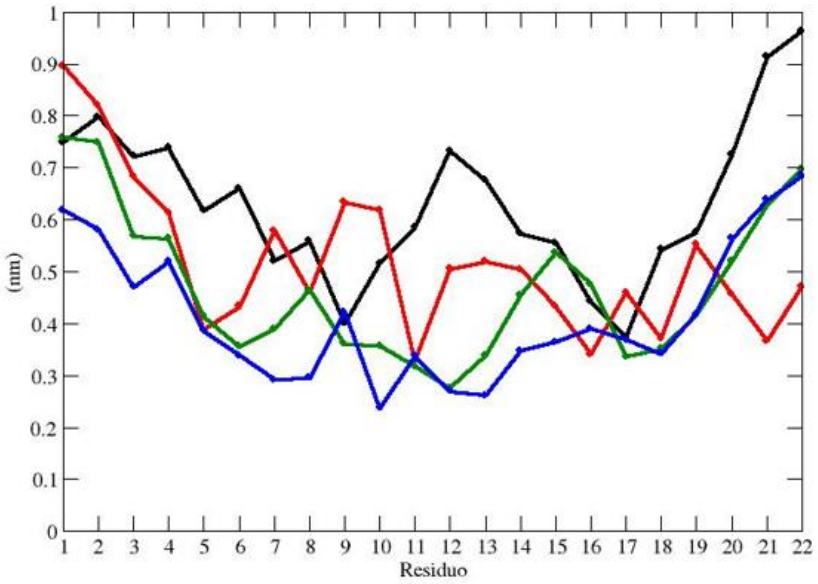


Figura 29. RMSF del péptido 1. Líneas: P1 línea negra, P1F línea roja, P1HA línea verde y P1FHA línea azul.

- Péptido 2

Los aminoácidos involucrados en la formación de puentes de hidrogeno corresponden en un 57% a treonina (residuo 2 y 19) y estos no presentan mucha variación con respecto a su RMSF (figura 29); de igual forma los aminoácidos con carga como el ácido glutámico (residuo 5) y lisina (residuo 18) no participan de manera importante en la formación de puentes de hidrogeno (tabla 22) y no presentan gran variación en su RMSF (figura 30).

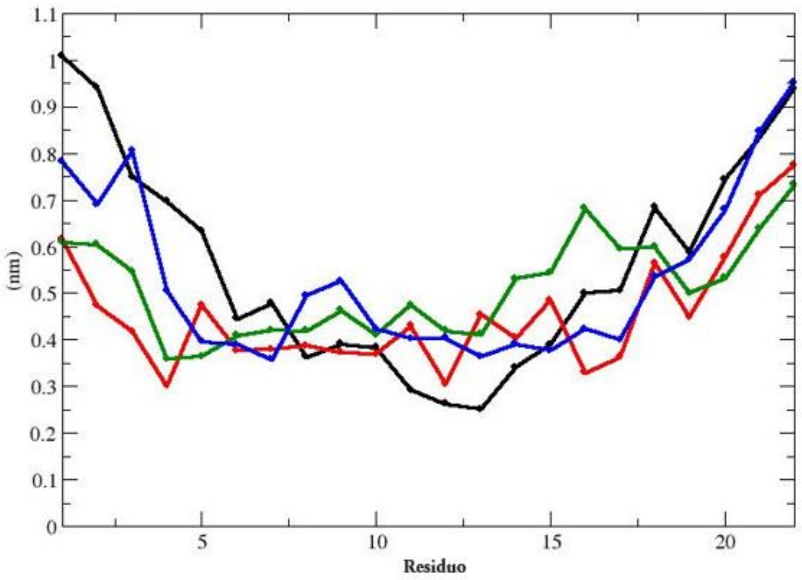


Figura 30. RMSF del péptido 2: Líneas P2 negra, P2F roja, P2HA verde y P2FHA azul.

- Péptido 3

Los aminoácidos que presentan mayor variación en el RMSF (figura 31) se encuentran involucrados en la formación de puentes de hidrógeno (tabla 22), en el P3 son principalmente aminoácidos sin carga, con excepción de la arginina n°16 que participa en 1/6 de los puentes de hidrógeno que se presentan con alta frecuencia; en el P3F participan en un 100% aminoácidos fosforilados de los cuales en el 90% participa SEP (residuos 6, 9, 18, 20 y 23); en los sistemas en presencia de HA la serina participa en 5/6 de los puentes de hidrogeno formados en los sistemas sin fosforilar; en la versión fosforilada SEP y TPO participan en 1/3 de los puentes de hidrógeno de alta frecuencia.

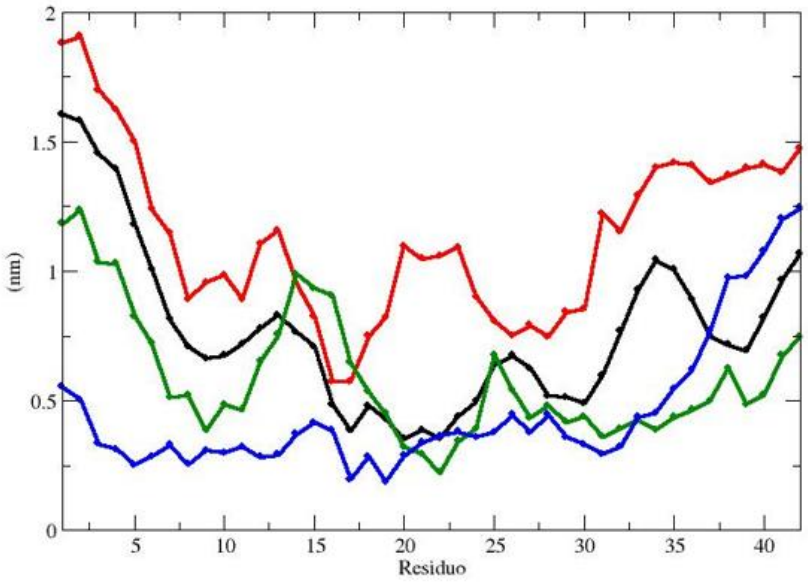


Figura 31. RMSF del péptido 3: Líneas P3 negra, P3F roja, P3HA verde y P3FHA azul.

- Péptido 4

Únicamente la lisina n° 61 (residuo n°12) se encuentra participando de manera frecuente en la formación de puentes de hidrógeno (tabla 22); como se observa en la figura 37 el RMSF de los aminoácidos con carga no presentan grandes variaciones con excepción de la lisina y valina en posición n° 55 y n° 64 (Figura 32, residuos 6 y 15 respectivamente).

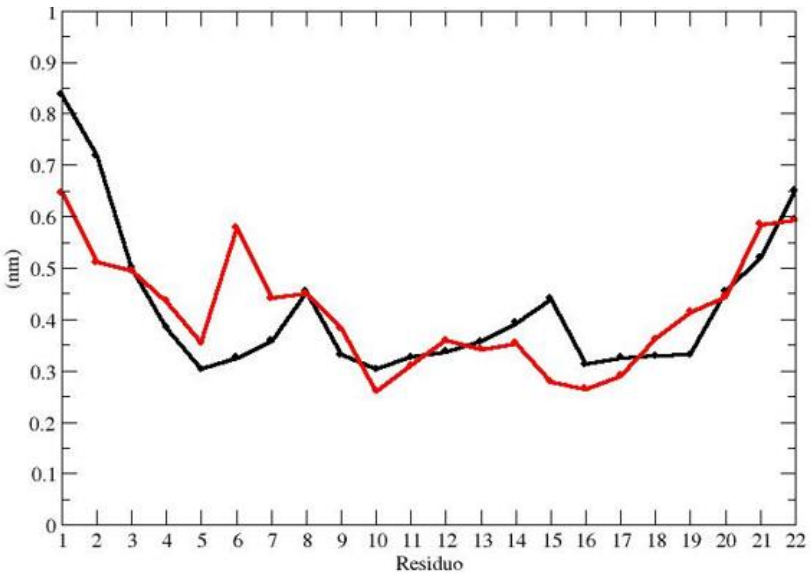


Figura 32. RMSF del péptido 4: Líneas P4 negra y P4 HA roja.

- Péptido 5

En el P5 participan de manera importante el aminoácido isoleucina n° 77 y 79 (residuo 8 y 10) en la formación de puentes de hidrógeno (tabla 22) y para el P5F los residuos de ácido glutámico n° 74 (residuo 5); sin embargo son pocos los puentes que se presentan con alta frecuencia esto se ve aumentado en el P5HA donde la participación de aminoácidos como alanina n°86 (residuo 17), valina n°75 (residuo 6), isoleucina n°77 y 79 (residuo 8 y 10) y la serina n° 84 (residuo 15) son importantes; sin embargo, en el P5FHA no se presentan puentes de hidrógeno con alta frecuencia. Si bien los aminoácidos antes mencionados no presentan carga se localizan cerca de aminoácidos con carga que no presentan grandes variaciones en su RMSF figura 33.

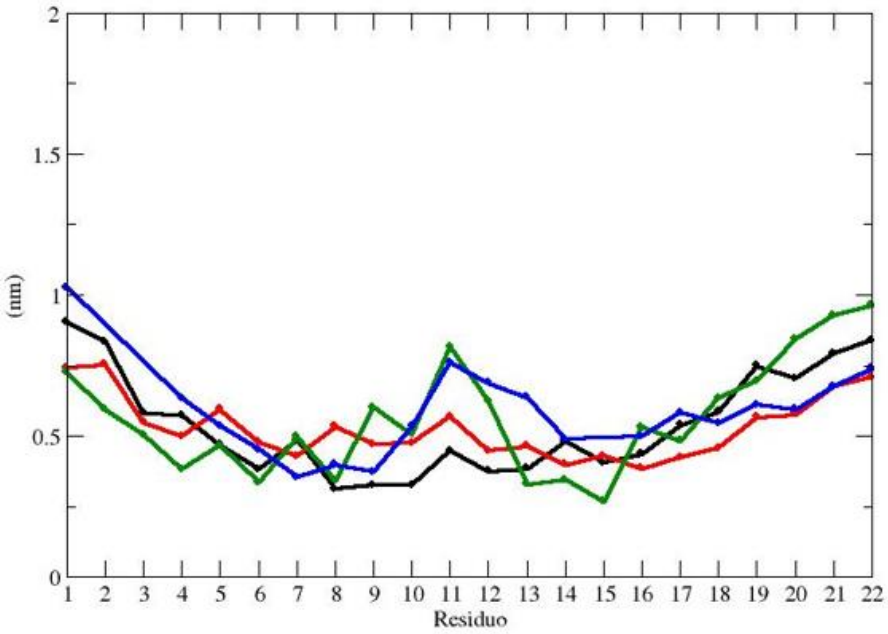


Figura 33. RMSF del péptido 5: Líenas P5 negra, P5F roja, P5HA verde y P5FHA azul.

DSSP

Todos los resultados de DSSP se resumen en la tabla 21.

- Péptido 1. El mayor porcentaje de estructura desordenada se presenta en los sistemas fosforilados, así mismo la presencia de estructura α-hélice es mínima para todos los sistemas con un porcentaje máximo de 0.13% en el sistema del P1; para el caso de estructura láminaβ, el porcentaje es mayor siendo de hasta 18.45% para P1HA.
- Péptido 2. El menor porcentaje de estructura desordenada se presenta en el sistema del péptido P2F con 79.99% y el porcentaje de estructura secundaria corresponde únicamente a láminaβ; los demás sistemas presentan por encima del 90% de estructura desordenada llegando hasta un 100% en el P2FHA.
- Péptido 3. El mayor porcentaje de estructura secundaria se presenta en las versiones sin fosforilar y el mayor porcentaje de estructura aleatoria en las versiones fosforiladas, de estas el sistema en presencia de HA es el que muestra mayor porcentaje con 98.51%.
- Péptido 4. El mayor porcentaje de estructura secundaria α-hélice se presenta en el P4 con 0.017% y el mayor porcentaje de estructura secundaria lámina-β se presenta en P4HA con 8.28% y el P4 presenta el mayor porcentaje de estructura aleatoria con 94.39%.
- Péptido 5. El P5HA presenta el menor porcentaje de estructura desordenada con 61.7% y el contra parte el P5FHA presenta el mayor porcentaje de estructura desordenada con 99.4%.

Gráficos de Ramachandran

- Péptido 1

La distribución de los ángulos de torsión por medio del gráfico de Ramachandran se muestra en la figura 34. Para todos los sistemas prevalece un valor promedio para ángulos correspondientes a estructura desordenada por arriba del 90% y el mayor porcentaje de estructura secundaria α-hélice (4.24%) se presenta en el péptido P1F; para estructura lámina β el mayor porcentaje (4.12%), se presenta en el sistema del péptido P1HA.

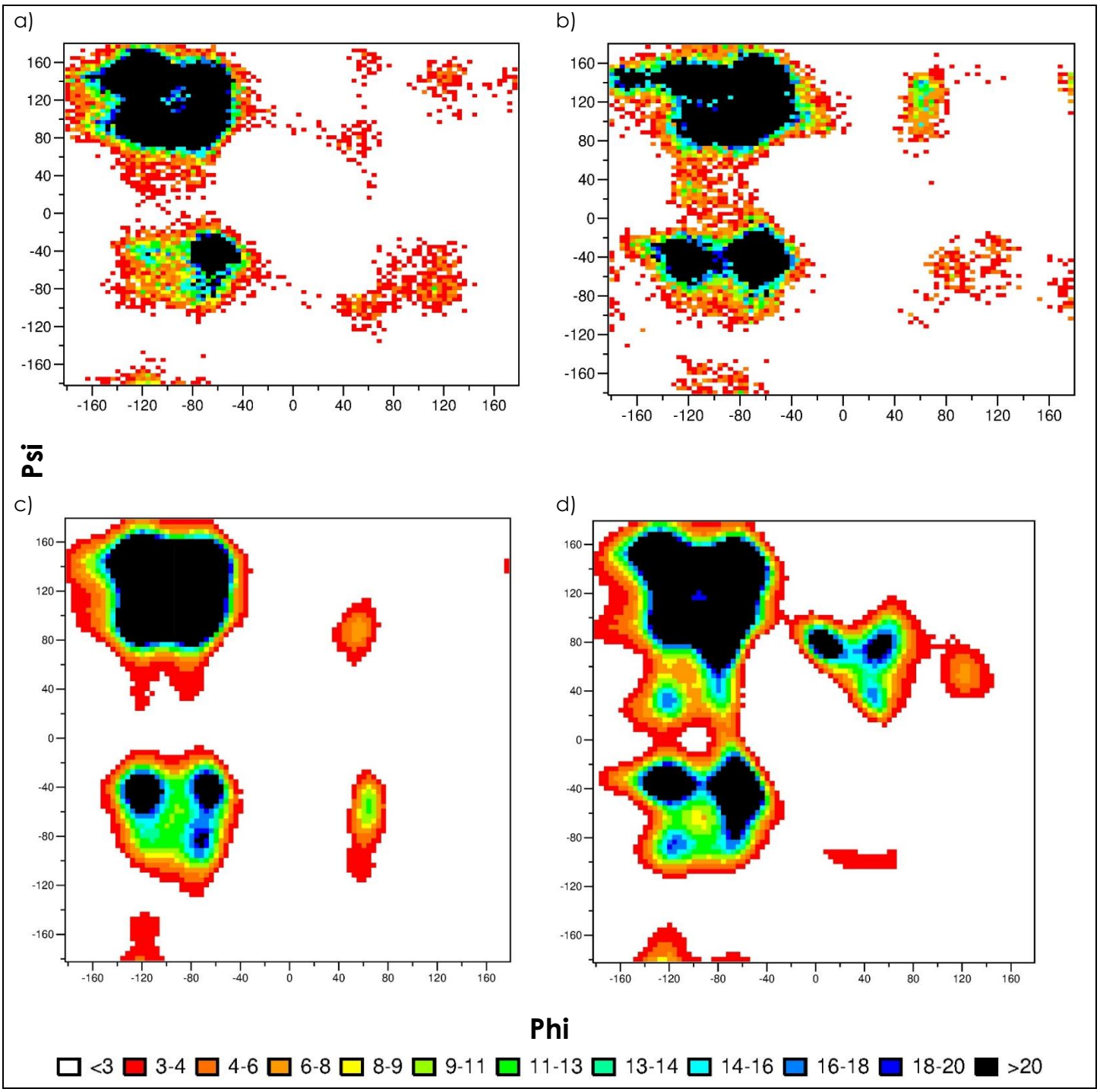


Figura 34. Gráficos de Ramachandran de los péptidos 1. a) P1, b) P1F, c) P1HA y d) P1FHA. La escala de colores indica el número de veces que se presentan los ángulos.

- Péptido 2

La distribución de los ángulos de torsión por medio del gráfico de Ramachandran se muestra en la figura 35. Para todos los sistemas prevalece un valor promedio para ángulos correspondientes a estructura desordenada por arriba del 90% y el mayor porcentaje de estructura secundaria α-hélice 4.08% se presenta en el P2HA; para estructura lámina β el mayor porcentaje 4.30%, se presenta en el P2F.

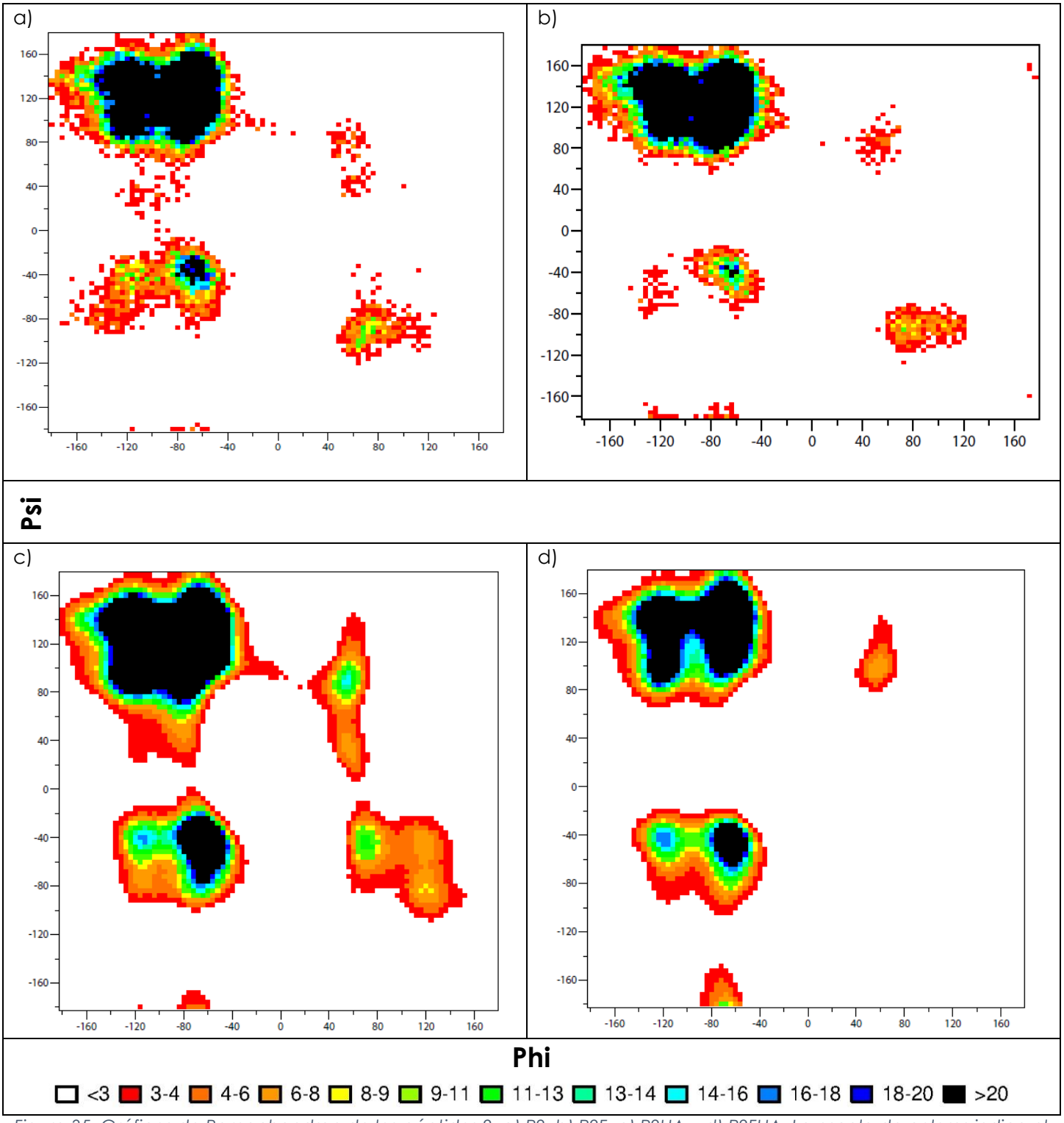


Figura 35. Gráficos de Ramachandran de los péptidos 2. a) P2, b) P2F, c) P2HA y d) P2FHA. La escala de colores indica el número de veces que se presentan los ángulos.

- Péptido 3

La distribución de los ángulos de torsión por medio del gráfico de Ramachandran se muestra en la figura 36. Para todos los sistemas prevalece un valor promedio para ángulos correspondientes a estructura desordenada por arriba del 90% y el mayor porcentaje de estructura secundaria tanto de estructura α-hélice 3.96% como de lámina β se presenta en el P3HA.

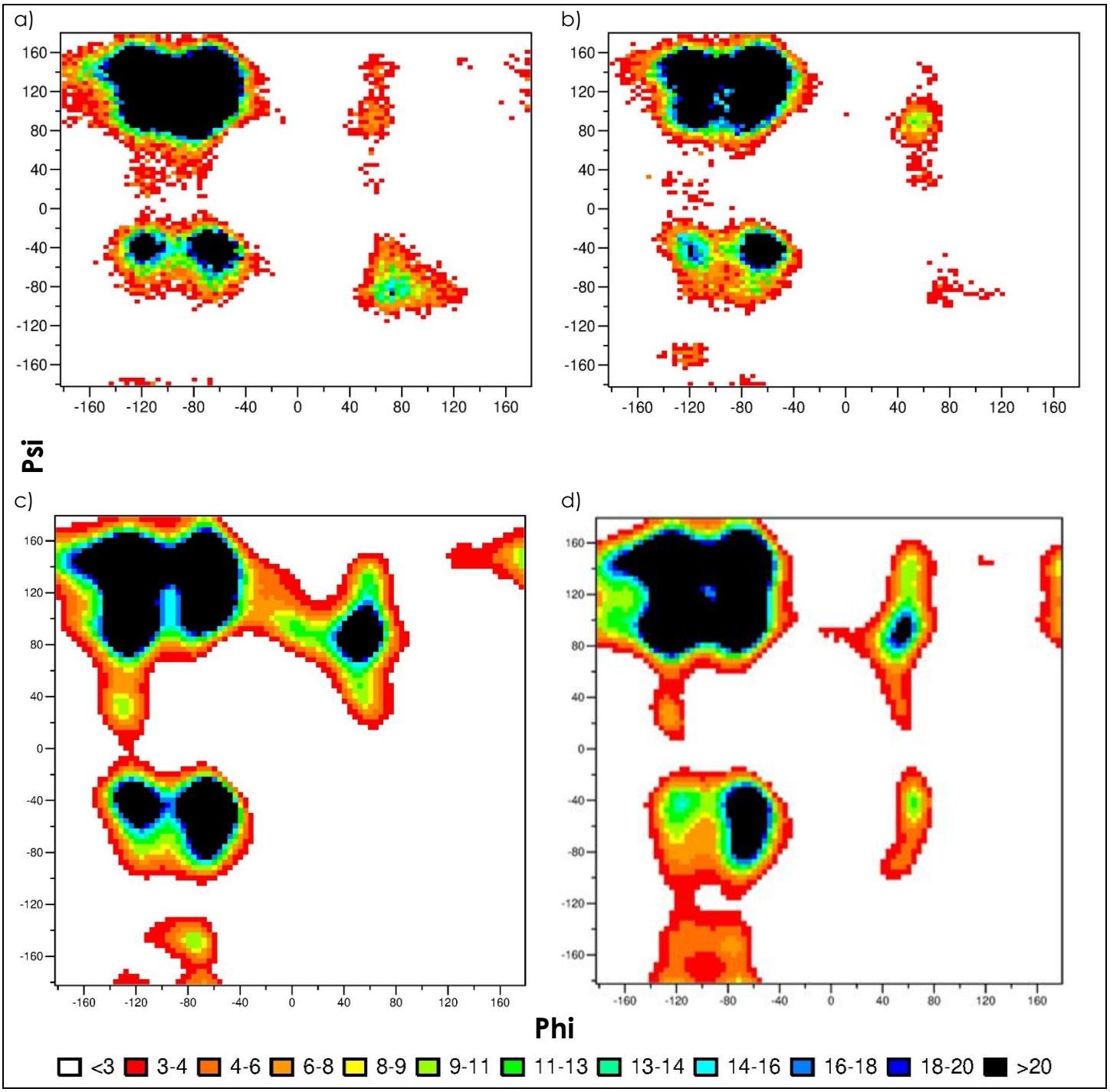


Figura 36. Gráficos de Ramachandran de los péptidos 3. a) P3, b) P3F, c) P3HA y d) P3FHA. La escala de colores indica el número de veces que se presentan los ángulos.

- Péptido 4

La distribución de los ángulos de torsión por medio del gráfico de Ramachandran se muestra en la figura 37, el menor porcentaje de estructura desordenada se presenta en el péptido P4HA con un 88.99% y coinciden con los resultados de DSSP donde de igual manera el menor porcentaje de estructura aleatoria se presenta en P4HA.

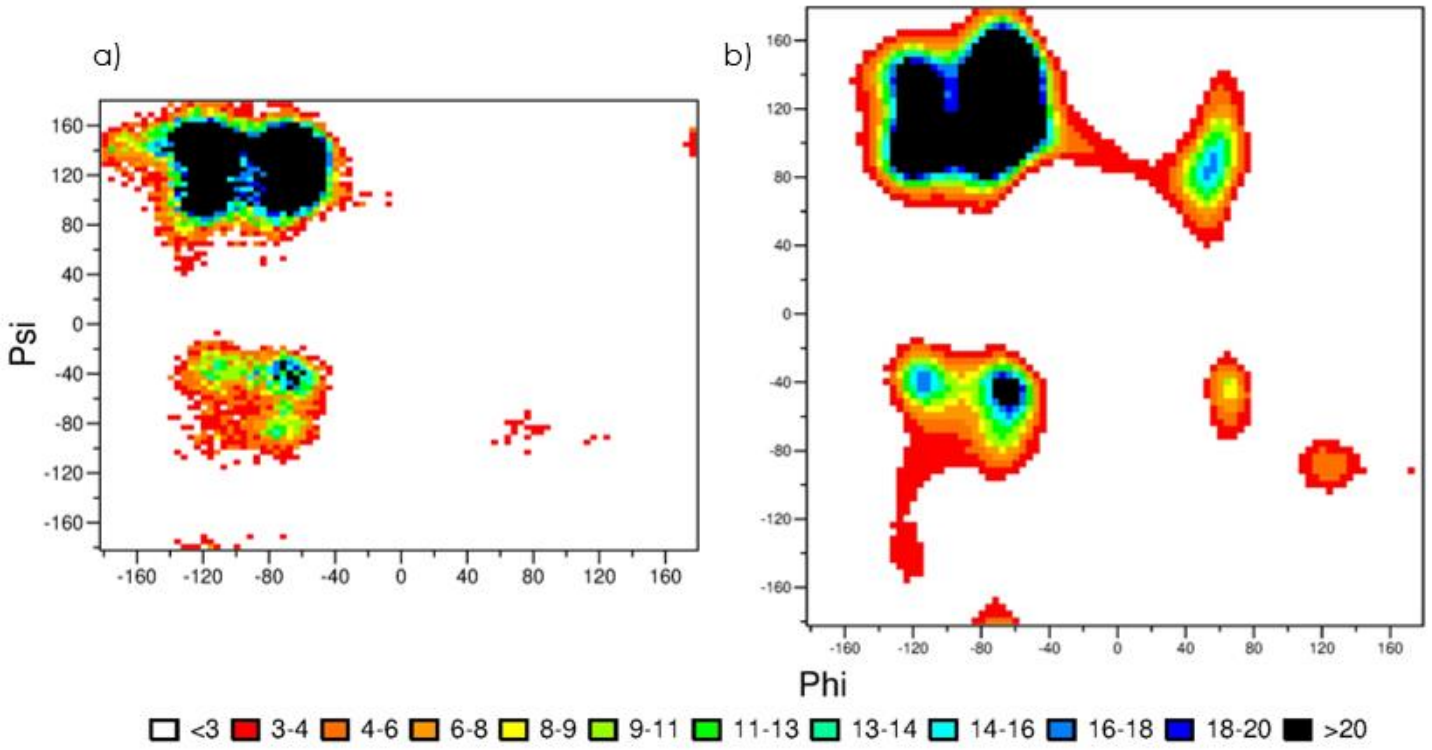


Figura 37. Gráficos de Ramachandran de los péptidos 4. a) P4 y b) P4HA. La escala de colores indica el número de veces que se presentan los ángulos.

- Péptido 5

La distribución de los ángulos de torsión por medio del gráfico de Ramachandran se muestran en la figura 38, todos los sistemas presentan por arriba del 90% de estructura desordenada, el mayor porcentaje de estructura secundaria α-hélice se presenta en P5F y el mayor porcentaje de estructura secundaria lámina β en P5HA.

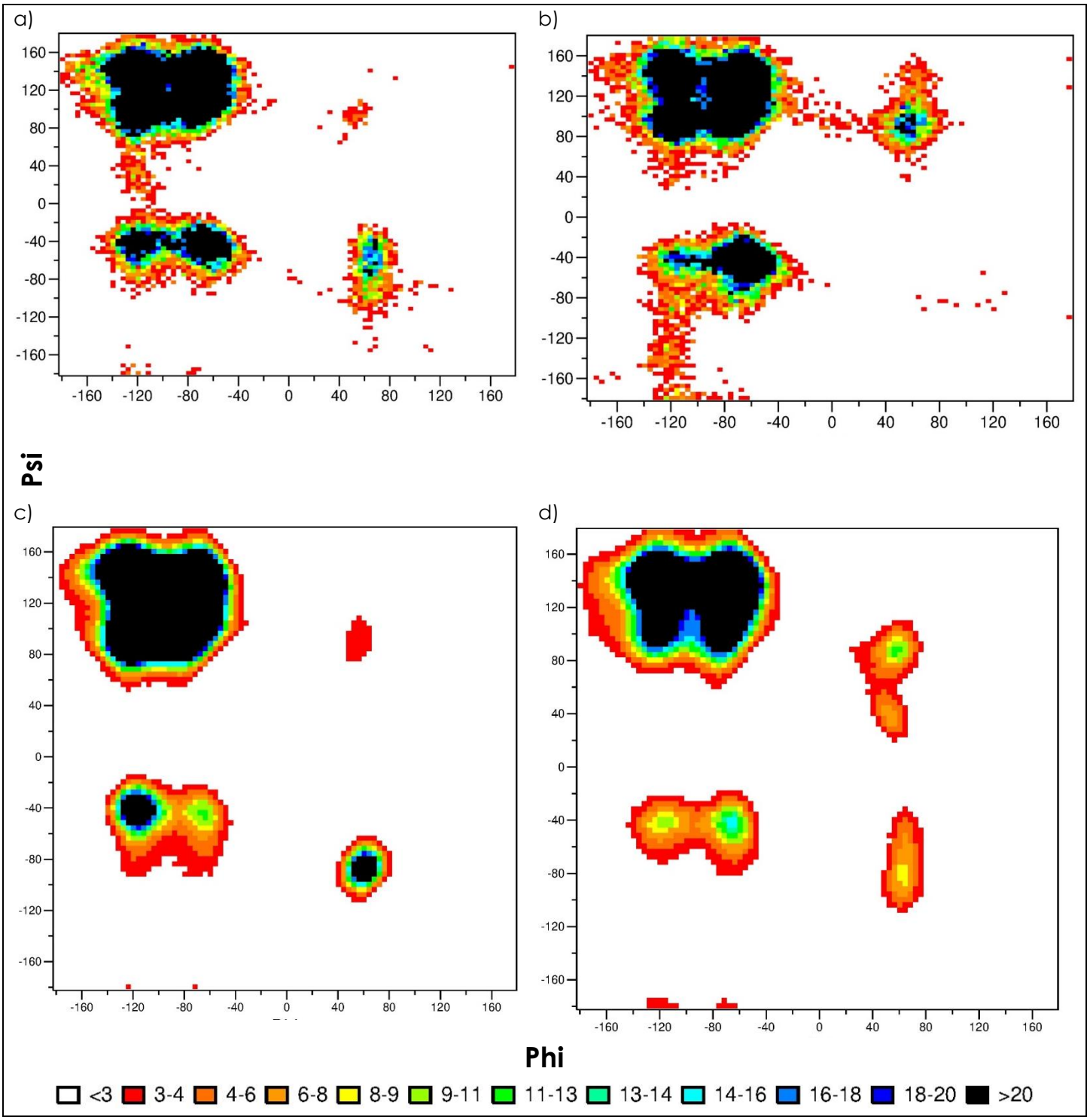


Figura 38. Gráficos de Ramachandran de los péptidos 5. a) P5, b) P5F, c) P5HA y d) P5FHA. La escala de colores indica el número de veces que se presentan los ángulos.

En la tabla 21 se muestran los resultados de estructura secundaria de todos los péptidos, en la columna de lado izquierdo se muestran los porcentajes de DSSP y en la columna de lado derecho los porcentajes correspondientes a los gráficos de Ramachandran.

Péptido	DSSP			Ramachandran		
	ESTRUCTURA					
	α %	β %	Coil %	α %	β %	Coil %
P1	0.13	14.97	84.90	3	3.34	93.66
P1F	0.09	2.82	97.09	4.24	1.44	94.32
P1HA	0.005	18.45	81.545	2.08	4.12	93.8
P1FHA	0.011	0.007	99.982	3.58	2	94.4
P2	0	2.52	97.48	2.26	4.2	93.54
P2F	0	20.01	79.99	1.42	4.30	94.28
P2HA	0.05	8.48	91.47	4.08	3.25	92.67
P2FHA	0	0	100	3.11	3.50	93.39
P3	0.19	12.12	87.69	2.88	3.35	93.77
P3F	0.008	3.21	96.78	3.79	3.20	93.01
P3HA	3.32	12.53	84.15	3.96	3.45	92.59
P3FHA	0.46	1.03	98.51	3.79	3.20	93.01
P4	0.017	5.59	94.39	1.81	4.01	94.18
P4HA	0.07	8.28	91.65	1.96	9.05	88.99
P5	1.36	11.59	87.05	3.2	3.25	93.55
P5F	0	3.86	96.14	5.88	2.34	91.78
P5HA	0	38.3	61.7	0.76	5.44	93.8
P5FHA	0.53	0.07	99.4	1.77	4.35	93.88

Tabla 21. Resultados de estructura secundaria de DSSP y gráficos de Ramachandran.

Puentes de hidrógeno

En la tabla 22 se muestran los resultados obtenidos por medio del programa HBonanza (Análisis de puentes de hidrógeno) (Durrant and McCammon, 2011); para la obtención de los puentes de hidrógeno formados durante las simulaciones de SDM. Se muestran: elementos químicos que forman el puente de hidrógeno (su numeración es de acuerdo al archivo pdb) y su tipo, el número de aminoácido (de acuerdo a su posición en la secuencia de CEMP1) y su tipo, y la última columna corresponde a la frecuencia con la que se presentan.

N° de Péptido	Índices atómicos		Aminoácido		Frecuencia
	N°	Tipo	N°	Tipo	
P1	135-134-68	H-N-O	15-15-8	ARG-ARG-SER	0.6641
	82-81-116	H-N-O	10-10-13	GLN-GLN-HIS	0.6457
	62-61-150	H-N-O	8-8-15	SER-SER-ARG	0.4710
P1F	150-149-173	HE-NE-O1P	15-15-17	ARG-ARG-SEP	0.9592
	156-155-175	1HH2-NH2-O3P	15-15-17	ARG-ARG-SEP	0.9132
	137-135-173	2HH1-NH1-O1P	14-14-17	ARG-ARG-SEP	0.8440
	140-138-174	2HH2-NH2-O2P	14-14-17	ARG-ARG-SEP	0.7872
	157-155-193	2HH2-NH2-O1P	15-15-19	ARG-ARG-SEP	0.5537
	91-90-66	H-N-O	10-10-7	GLN-GLN-ASP	0.5161
	199-198-193	H-N-O1P	20-20-19	ASN-ASN-SEP	0.5085
	206-204-175	2HD2-NP2-O3P	20-20-17	ASN-ASN-SEP	0.4802
	154-152-194	2HH1-NH1-O2P	15-15-19	ARG-ARG-SEP	0.4598
	59-58-125	H-N-O	7-7-13	ASP-ASP-HIS	0.4230
P1HA	82-81-174	H-N-O	10-10-18	GLN-GLN-THR	0.657
	100-99-150	H-N-O	12-12-15	GLY-GLY-ARG	0.443
	184-183-68	H-N-O	20-20-8	ASN-ASN-SER	0.431
P1FHA	32-31-194	HG-OG-O2P	5-5-19	SER-SER-SEP	0.5821
	183-182-175	HG1-OG1-O3P	18-18-17	THR-THR-SEP	0.5137
	123-122-174	HE2-NE2-O2P	13-13-17	HIS-HIS-SEP	0.5125
	28-27-194	H-N-O2P	5-5-19	SER-SER-SEP	0.5084
	156-155-173	1HH2-NH2-O1P	15-15-17	ARG-ARG-SEP	0.4727
	179-178-175	H-N-O3P	18-18-17	THR-THR-SEP	0.4231
P2	153-152-12	H-N-O	39-39-21	ALA-ALA-THR	0.4686
	91-90-151	H-N-O	31-31-38	LEU-LEU-THR	0.4078
P2F	17-16-10	H-N-O1P	22-22-21	SEP-SEP-TPO	0.6257
	80-79-142	H-N-O	29-29-36	LEU-LEU-GLY	0.5857
	156-155-71	H-N-O	38-38-27	THR-THR-THR	0.5433
	139-138-87	H-N-O	36-36-29	GLY-GLY-LEU	0.4762
	64-63-163	H-N-O	27-27-38	THR-THR-THR	0.4662
P2HA	0	0	0	0	0
P2FHA	0	0	0	0	0
P3	273-272-193	H-N-O	30-30-20	SER-SER-ASN	0.7161
	184-183-279	H-N-O	20-20-30	ASN-ASN-SER	0.6593
	204-203-262	H-N-O	22-22-28	SER-SER-CYS	0.6085
	135-134-80	H-N-O	15-15-9	ARG-ARG-GLN	0.5561
	281-280-320	H-N-O	31-31-36	LEU-LEU-GLY	0.5217
	152-151-87	H-N-OE1	16-16-10	CYS-CYS-GLN	0.4270
P3F	150-149-173	HE-NE-O1P	15-15-17	ARG-ARG-SEP	0.8656
	137-135-173	2HH1-NH1-O1P	14-14-17	ARG-ARG-SEP	0.8376
	233-232-228	H-N-O2P	23-23-22	ALA-ALA-SEP	0.7504
	156-155-175	1HH2-NH2-O3P	15-15-17	ARG-ARG-SEP	0.7373
	199-198-194	H-N-O2P	20-20-19	ASN-ASN-SEP	0.7221
	140-138-174	2HH2-NH2-O2P	14-14-17	ARG-ARG-SEP	0.6957
	205-204-194	1HD2-ND2-O2P	20-20-19	ASN-ASN-SEP	0.5729
	86-85-74	1HE2-NE2-O2P	9-9-8	GLN-GLN-SEP	0.5213
	99-97-43	2HE2-NE2-O3P	10-10-5	GLN-GLN-SEP	0.4922

	256-254-216	2HD2-ND2-O2P	25-25-21	ASN-ASN-TPO	0.4818
	98-97-73	1HE2-NE2-O1P	10-10-8	GLN-GLN-SEP	0.4742
P3HA	62-61-193	H-N-O	8-8-20	SER-SER-ASN	0.704
	184-183-68	H-N-O	20-20-8	ASN-ASN-SER	0.638
	82-81-179	H-N-O	10-10-19	GLN-GLN-SER	0.581
	44-43-210	H-N-O	6-6-22	THR-THR-SER	0.572
	228-227-17	H-N-O	25-25-2	ASN-ASN-GLY	0.535
	334-333-308	H-N-O	38-38-34	THR-THR-SER	0.504
P3FHA	161-160-381	H-N-O	16-16-40	CYS-CYS-PRO	0.571
	222-221-215	H-N-O1P	22-22-21	SEP-SEP-TPO	0.485
	144-143-125	H-N-O	15-15-13	ARG-ARG-HIS	0.483
P4	84-83-152	H-N-O	61-61-69	LYS-LYS-PRO	0.8064
	96-95-15	H-N-O	62-62-52	ALA-ALA-GLN	0.4814
P4HA	125-124-82	H-N-O	66-66-60	ILE-ILE-VAL	0.654
	70-69-38	H-N-O	59-59-54	ALA-ALA-PRO	0.472
	24-23-152	H-N-O	53-53-69	LEU-LEU-PRO	0.442
	52-51-74	H-N-O	56-56-59	GLY-GLY-ALA	0.4018
P5	69-68-50	H-N-O	77-77-75	ILE-ILE-VAL	0.5473
	91-90-20	H-N-O	79-79-89	ILE-ILE-GLY	0.4466
P5F	50-49-46	H-N-OE2	75-75-74	VAL-VAL-GLU	0.6481
	89-88-46	HD1-ND7-OE2	78-78-74	HIS-HIS-GLU	0.5637
P5HA	181-180-5	H-N-O	86-86-75	ALA-ALA-VAL	0.722
	91-90-141	H-N-O	79-79-82	ILE-ILE-LEU	0.679
	69-68-158	H-N-O	77-77-84	ILE-ILE-SER	0.674
	44-43-185	H-N-O	75-75-86	VAL-VAL-ALA	0.624
	152-151-76	H-N-O	84-84-77	SER-SER-ILE	0.601
P5FHA	0	0	0	0	0

Tabla 22. Puentes de hidrógeno formados durante SDM de péptidos..

Discusión

CEMP1 es una proteína con un peso molecular teórico calculado de 26kDa y con modificaciones postraduccionales aumenta a 50 kDa (Álvarez-Pérez et al., 2006). Existen estudios de caracterización físico-química de proteína CEMP1 recombinante humana hrCEMP1(Villarreal-Ramírez et al., 2009) y CEMP1 de plásmidos de la levadura *Pichia pastoris* (Romo-Arévalo et al., 2016), dichos estudios han reportado diferentes contenidos de estructura secundaria.

Villarreal-Ramírez et al. en el 2009, reportaron un contenido de estructura secundaria de 10% de α-hélice, 38.2% de lámina β y un 35% de estructura aleatoria; estos datos se aproximan a los resultados obtenidos por medio de la predicción del estado desordenado de CEMP1, los cuales nos indican un 47.7% de estructura desordenada. Romo-Arévalo et al. en el 2016, reportaron que la estructura secundaria de CEMP1 está compuesta por un 28,6% de α-hélice, un 9,9% de lámina β y un 61,5% de estructura aleatoria; el contenido de estructutura aleatoria se asemeja al porcentaje promedio de los resultados de los programas de predicción de estructura secundaria en estructura aleatoria de 64.6%.

La diferencia en el contenido de estructura secundaria reportado experimentalmente de los autores antes mencionados, es debido a las modificaciones postraduccionales que cambian la estructura secundaria de CEMP1. Romo-Arévalo et al., utilizaron la levadura *Pichia pastoris* para la expresión de CEMP1, este sistema de expresión permite que se lleven a cabo como modificaciones postraduccionales las N-glicosilaciones (Ahmad et al., 2014; Cregg et al., 2000; Daly and Hearn, 2005; Damasceno et al., 2012; Macauley-Patrick et al., 2005), como resultado el peso molecular de CEMP1 aumenta de 25.9kDa a 28.77kDa (Romo-Arévalo et al., 2016). Por otra parte, Villarreal-Ramírez et al., utilizaron fibroblastos gingivales humanos para la expresión de la proteína CEMP1 lo cual permite además de glicosilaciones, las fosforilaciones como modificaciones postraduccionales resultando en un peso molecular para CEMP1 de 50kDa (Villarreal-Ramírez et al., 2009). Como sabemos las modificaciones postraduccionales participan en la adquisición de estructura secundaria (Alberts Bruce et al., 2016) y por ello las diferencias en los porcentajes de estructura secundaria reportados. Además, los diferentes sistemas para la expresión de proteínas utilizados presentan ventajas y desventajas, entre ellas se encuentran el posibilitar o no las modificaciones postraduccionales y de igual forma el tipo de modificación que puede realizar el sistema de expresión (Dingermann, 2008; Gomes et al., 2016; Rai and Padh, 2001; Yin et al., 2007).

Con respecto a los distintos resultados obtenidos de los predictores de estructura utilizados, la diferencia radica en la manera en que funciona cada programa. El algoritmo de PONDR utiliza las características de secuencias desordenadas conocidas (composición de aminoácidos, la tendencia de cada aminoácido de promover el desorden estructural, baja complejidad de secuencia, alta variabilidad de secuencia y bajo predicción de contenido de estructura secundaria), para la predicción del desorden estructural (Ferron et al., 2006). Por otro lado los predictores de estructura secundaria, están basados en métodos probabilísticos o redes neuronales artificiales que usan la información de bases de datos de la estructura de proteínas globulares resueltas experimentalmente (Cai et al., 2003, 2002; Chou, Peter Y. and Fasman, Gereald D., 1974; Chou and Fasman, 1977; Garnier, Jean et al., 1996; Geourjon and Deléage, 1995; Kumar, 2013), los cuales presentan diferencias estructurales con proteínas intrínsecamente desordenadas.

La proteína CEMP1 se ha reportado como una proteína teóricamente alcalina con un punto isoeléctrico de 9,73 (Arzate et al., 2015). Nuestros cálculos de la secuencia completa de

CEMP1 y sin modificaciones postraduccionales, indican una carga neta positiva de +13, correspondiente a una proteína alcalina. Sin embargo, el número de fosforilaciones en la secuencia de CEMP1, la convierten en una proteína altamente negativa con una carga neta de -47. El mismo fenómeno se presentan en los péptidos, cuando se encuentran sin fosforlaciones la carga neta es positiva en su mayoria o neutra y al ser fosforilados su carga neta es negativa.. Debido a que CEMP1 no cuenta con un homólogo en el genoma, comparamos nuestros resultados con lo reportado para otras proteínas no colágenas asociadas al proceso de biomineralización como la osteopontina (OPN), la sialoproteína ósea (BSP), la proteína de matriz de dentina 1 (DMP1) y la sialofosfoproteína de dentina (DSPP), , (George and Veis, 2008; Qin et al., 2004). Todas las proteínas antes mencionadas son miembros de la familia de proteínas SIBLING. Las proteínas SIBLING presentan características en común con la proteína CEMP1, dentro de las cuales se incluyen afinidad a HA, abundancia de residuos con carga eléctrica, presentan modificaciones postraduccionales como la fosforilación y glicosilación, y altos porcentajes de estructura desordenada (Goldberg et al., 1996; Grzybowska, 2018; Qin et al., 2004; Wojtas et al., 2012).

En los péptidos 1 (P1, P1F, P1HA, P1FHA) observamos que la carga neta no afecta de manera importante al Rg, la variación registrada del Rg entre los diferentes sistemas se encuentra en un rango de 0.1nm. En estructura de proteínas no es una diferencia importante. Sin embargo, la diferencia en la carga que presenta el péptido afecta la cinética de unión a HA en los sistemas P1HA y P1FHA, como se puede observar por el rápido cambio en la distancia existente entre los péptidos y la hojuela, así como por la diferencia entre las distancias registradas, para P1HA la diferencia entre la distancia inicial y final es de 2.026nm, mientras que para P1FHA la diferencia es de 2.837nm, lo cual significa que se aproxima 0.8nm más a la hojuela de HA, cabe mencionar que este fenómeno se presentó en todos los péptidos exceptuando a P4HA que no cuenta con una versión fosforilada.

P1 presento un Rg de 0.7nm, siendo la versión más compacta de todos los péptidos 1. Los aminoácidos de serina, glutamina, histidina y arginina en las posiciones 8, 10,13 y 15 respectivamente, se encuentran en un estado desordenado (acuerdo a la predicción de desorden) y presentan mayor variación en el RMSF. Además, estos aminoácidos participan en la formación de 3 puentes de hidrógeno de P1 y pueden estar asociados a la formación de estructura secundaria de un 15.1% en el análisis de DSSP. La versión fosforilada P1F presento un Rg de 0.8nm y presenta 10 puentes de hidrógeno. Un aumento en comparación de P1. Los aminoácidos que participan en la formación de puentes de hidrógeno de manera importante son la arginina n°14 y 15, y la fosfoserina n°17 y 19. Sin embargo, los resultados de DSSP indican un alto porcentaje de estructura aleatoria con 97.09%. Determinamos que debido a la alta reactividad de los aminoácidos fosforilados es que se presenta un mayor número de puentes de hidrogeno, sin embargo estos no se forman de una manera estable por lo cual se presentan bajos porcentajes de estructura secundaria en los resultados de DSSP.

Para el P1HA con un Rg de 0.8nm presenta únicamente 3 puentes de hidrógeno en los cuales participan un 66% de aminoácidos en estado desordenado como la serina, glutamina, glicina y arginina en posiciones 8, 10, 12 y 15 respectivamente; los resultados de DSSP indican que es la versión que presenta el menor porcentaje de estructura aleatoria con un 81.5% y un aumento del porcentaje de estructura secundaria en láminas β con 18.45%. En el caso P1FHA con un Rg de 0.8nm, presenta 6 puentes de hidrógeno. Sin embargo los resultados de DSSP indican que es la versión del péptido 1 con mayor porcentaje de estructura aleatoria con un 99.9%. Comparado con los resultados de Ramachandran existe una

diferencia de 5.58% de estructura aleatoria. Sin embargo, ambos resultados nos indican que P1FHA es el péptido con mayor porcentaje de estructura aleatoria de todos los péptidos 1.

Si bien la literatura reporta la importancia y participación de las fosforilaciones para la adquisición de estructura secundaria, en el caso de P1 y P1F podemos observar un decremento del porcentaje de estructura secundaria, este mismo suceso ocurrió en presencia de HA, es decir observamos un aumento de estructura secundaria en P1HA y la disminución del mismo en P1FHA, por lo que postulamos que si bien las fosforilaciones hacen más eficiente la unión a HA, una vez unido el péptido la carga negativa proporcionada por las fosforilaciones se neutraliza de manera que no hay fuerzas electrostáticas que induzcan la formación de puentes de hidrógeno generando la permanencia de un estado desordenado. Exceptuando a el P4HA, observamos este fenómeno en todos los péptidos en presencia de HA, es decir una disminución de los porcentaje de estructura secundaria de DSSP de las versiones sin fosforilar a las versiones fosforiladas.

Además observamos un aumento del 3.48% de estructura secundaria de DSSP entre P1 y P1HA, de igual forma este aumento de estructura se presentó en todos los sistemas lo que indica una asociación entre la adquisición de estructura secundaria y la presencia de HA.

Los péptidos 2 (P2, P2F, P2HA, P2FHA) presentan una carga negativa -1 para las versiones sin fosforilar y -9 en las versiones fosforiladas. De acuerdo a los resultados de predicción de estructura desordenada de P2 la treonina en posición 21 se encuentra en estructura secundaria y la treonina 38 en estado desordenado, ambas participan en la formación de 2 puentes de hidrógeno y presentan variación en el RMSF y un Rg de 0.7nm. En P2F aumenta el número de puentes de hidrógeno a 5, donde la fosfoserina n° 22 y la fosfotreonina n° 21 en estado ordenado, y la glicina, leucina y treonina en posiciones 36, 29, 27 y 38 es estado desordenado son importantes. Los resultados de DSSP indican que presenta el menor porcentaje de estructura aleatoria con 79.99% y un aumento de estructura secundaria lámina β con 20.01%. Observamos un aumento del 17.49% de estructura secundaria lámina β en los resultados de DSSP entre P2 y P2F. El menor Rg de entre todos los péptidos estudiados, se presentó en P2HA con 0.04nm, pero no presenta puentes de hidrógeno con alta frecuencia (se presenten con una frecuencia por arriba del 0.4 tabla 21) y sus resultados de DSSP muestran un 91.47% estructura aleatoria. El mayor Rg se presentó en el P2FHA con 1.07nm, de igual manera no presenta puentes de hidrógeno, los resultados de DSSP indican un 100% de estructura aleatoria y los de Ramachandran un 93.39%.

Para el péptido n°3 observamos que al igual que el péptido n°1 la carga no afecta el Rg, el rango de diferencia entre los diferentes sistemas es de hasta 0.8nm sin embargo, considerando que este péptido está conformado por el doble de aminoácidos (40 aminoácidos) presenta altos grados de compactación similares al de los demás péptidos (1.0nm a 1.8nm). Al igual que en otros sistemas observamos en las versiones sin HA el aumento en la formación de puentes de hidrógeno en las versión fosforilada. Los resultados de DSSP indican un porcentaje de estructura aleatoria de 87.69% para P3 y disminuye este porcentaje a 84.1% en presencia de HA, presenta menor puentes de hidrógeno y con menor frecuencia lo que indica que el aumento de estructura secundaria puede estar asociado a la interacción con HA. En cuanto a los resultados de estructura secundaria de los gráficos de Ramachandran indican porcentajes por arriba del 90% para todos los sistemas.

El P4 presenta mayor Rg, menor grado de compactación debido a la poca cantidad de puentes de hidrógeno que se forman, de acuerdo a la predicción de desorden, un 25% de los aminoácidos involucrados en la formación de los puentes de hidrógeno se encuentran

en un estado desordenado. Para el caso de P4HA observamos el cambio en el Rg a un mayor grado de compactación, asociado a un aumento en la formación de puentes de hidrógeno que involucran un 66.6% de aminoácidos en estado desordenado; a pensar de que este péptido no cuenta con posibles sitios de fosforilación la distribución de aminoácidos cargados positivamente como la lisina en las posiciones 5 y 11, y la histidina en la posición 20, así como aminoácidos cargados negativamente como el ácido glutámico en la posición 13 permiten tanto la formación de puentes de hidrógeno antes y después de su unión a la HA, cabe mencionar que es por medio de este aminoácido cargado negativamente que se genera la unión del péptido con HA; de igual manera estos resultados correlacionan con los porcentajes calculados de estructura secundaria por medio de DSSP y gráficos de Ramachandran, en los cuales podemos observar que los menores porcentajes de estructura secundaria se presentan en P4.

En el péptido n° 5 la carga que presentan los péptidos sin fosforilar es de +3 y fosforilados de -3 sin embargo, esto no afecta considerablemente al Rg que cambia un rango de 0.1nm entre las diferentes versiones. La versión más compacta se presenta en P5 con Rg 0.8nm, con únicamente dos puentes de hidrógeno de alta frecuencia, los cuales aumentan en la versión P5HA, esto correlaciona con los resultados de DSSP los cuales indican un 87.05% de estructura aleatoria para P5 y un 61.7% para P5HA. El P5FHA con un Rg de 0.8nm no presenta puentes de hidrógeno con alta frecuencia por lo que sus resultados de DSSP indican alto porcentaje de estructura aleatoria con 99.4%. En cuanto a los resultados de estructura secundaria de los gráficos de Ramachandran indican porcentajes por arriba del 90% para todos los sistemas.

Con excepción del péptido n° 2, notamos que la carga no afecta de manera importante al radio de giro, los rangos de diferencia entre el estado desfosforilado y fosforilado oscila entre los 0.1 nm y 0.2 nm. Sin embargo cuando los péptidos se encuentran en presencia de HA se observa una alteración en el Rg. Por ejemplo P2HA que presenta un Rg de 0.048nm y su versión fosforilada P2FHA presenta un Rg de 1.07nm. También P4 presenta un Rg de 0.8nm y en presencia de HA (versión P4HA) su Rg es de 0.08nm.

Determinamos a P2HA como el péptido con menor Rg debido a la formación de puentes de hidrógeno. De acuerdo con los resultados de predicción del estado desordenado la mayoría de los aminoácidos asociados a la formación de estos enlaces se encuentran desordenados. El péptido interactúa con la HA a través del ácido glutámico n° 24, disminuyendo la formación de puentes de hidrógeno; una vez que se mantiene la interacción por medio de este aminoácido se retoma la formación de dichos enlaces. En caso de P2FHA presenta un mayor Rg, un menor grado de compactación, observamos que los aminoácidos como serina y treonina fosforilados interactúan rápidamente (en un tiempo de 3ps) con la HA y en consecuencia de la unión péptido-HA existe una nula formación de puentes de hidrógeno. Estos resultados concuerdan con los porcentajes obtenidos de estructura desordenada por medio de DSSP y los gráficos de Ramachandran.

Los cambios en el radio de giro se encuentran asociados a la formación de puentes de hidrógeno; que a su vez se presentan con mayor frecuencia de manera inicial en la SDM cuando los péptidos no se encuentran en un estado fosforilado. Pero no están los péptidos P1F, P3F, P2F entre los que más puentes de hidrógeno tienen. Para los péptidos fosforilados la unión con HA es prácticamente en cuestión de pico segundos, en algunos casos posterior a la interacción péptido-HA es que se da la formación de puentes de hidrógeno, una vez que se presenta la interacción con la HA, los puentes disminuyen en frecuencia para en algunos casos retomar la formación de puentes posterior a este evento, estos resultados

refuerzan estudios que reportan la capacidad de la HA de inducir el plegamiento intramolecular de péptidos (Capriotti et al., 2007).

La proteína CEMP1 posee una alta afinidad por la HA, sin embargo el mecanismo de interacción específica entre CEMP1 y la HA no es completamente comprendido; los resultados obtenidos por medio de las SDM indican que la presencia de aminoácidos con carga positiva (H, R y K), negativa (D y E) y aún más importante aminoácidos fosforilados, son importantes para determinar en qué superficie de la hojuela de HA se unirá; dado que en las superficies de la hojuela se presentan expuestos a diferentes átomos, por lo cual se postula que el mecanismo de interacción entre los péptidos derivados de CEMP1 y HA es una interacción de tipo electrostática.

La presencia de motivos de aminoácidos con carga positiva, como es el motivo HRR (posición 13-15 en secuencia de aminoácidos de CEMP1), en los péptidos P1HA y P3HA, permite la unión de los péptidos hacia la superficie de la hojuela que presenta expuestos los grupos fosfato. De igual forma para las versiones fosforiladas se observó una rápida y cercana interacción entre los péptidos y HA. Debido a las fosforilaciones consideradas como sitios de unión para los iones de calcio (Villarreal-Ramírez et al., 2017), se generan tripletes cargados negativamente TDS (posición 7-9) en P1F y P3F que son de vital importancia para hacer más eficiente la unión, dirigiendo el péptido hacia la cara de la hojuela de HA que presenta expuestos átomos de calcio; estos resultados concuerdan con otros estudios que reportan la importancia del estado fosforilado en proteínas relacionadas con el proceso de biomineralización como las proteínas DPP, OPN y DMP1(George and Veis, 2008; He et al., 2005; Narayanan et al., 2003; Qin et al., 2004), así como de la importancia de las regiones poliácidas que promueven la unión a los biominerales al interactuar con iones inorgánicos y superficies minerales, estas regiones poliácidas son comunes en muchas proteínas endógenas de unión a hueso, como la estaterina, la osteonectina, la sialoproteína ósea y la osteopontina, por lo que se considera que dirigen el reconocimiento del biomineral (Gilbert et al., 2000; Sawyer et al., 2005)

Las secuencias de aminoácidos ácidas se han utilizado para lograr anclar ionicamente péptidos bioactivos en una orientación favorable sobre superficies inorgánicas (Sawyer et al., 2005). En estas se incluyen motivos de unión a HA como es el caso del motivo N15 de la estaterina que reconoce la HA a través de la secuencia pentapeptídica ácida N-terminal, que contiene dos serinas fosforiladas (Gilbert et al., 2003; Makrodimitris et al., 2007), el motivo EEEEEEEPRGDT presente en la sialoproteina osea donde la secuencia consecutiva de ácido glutámico media la unión con HA (Fujisawa et al., 1997; Itoh et al., 2002), mientras que el motivo RGD media la unión celular por medio de integrinas (Ruoslahti, 2003). En el 2000 Cross et al. mencionan diferentes motivos de unión caracterizados por multiples residuos de serinas fosforilados como el motivo Ser(p)-Ser(p)-Ser(p)-Glu-Glu presente en la fosfoproteína de dentina, el motivo Ser(p)-Ser(p)-Gly-Ser(p)-Ser(p)-Glu-Glu presente en la osteopontina, el motivo Asp-Ser(p)-Ser(p)-Glu-Glu presente en la estaterina entre otros que estabilizan el fosfato de calcio amorfo y se unen con alta afinidad a fases de fosfato de calcio cristalino (Cross et al., 2000).

De manera general observamos una tendencia al aumento de estructura secundaria entre los péptidos sin fosforilar y fosforilados en ausencia de HA (con excepción de P3 y P5). Si bien las fosforilaciones ejercen un papel importante en la adquisición de estructura secundaria (Villarreal-Ramírez et al., 2017) cuando estas participan en la unión entre el péptido y HA existe una disminución en los porcentajes de estructura secundaria debido a la participación de dichos aminoácidos fosforilados en la unión a HA en lugar de su

participación en la formación de puentes de hidrógeno estos resultados concuerdan con los cambios en el Rg de los péptidos en los cuales observamos un mayor acercamiento en las versiones fosforiladas. Estos resultados concuerdan con lo reportado por He et al. en el 2005 donde demuestran que la DPP no fosforilada tiene una capacidad de unión al calcio menor que la forma fosforilada e induce la formación de fosfato de calcio amorfo, mientras que la forma fosforilada promueve la formación de cristales de apatita.

Actualmente se han reportado diversas proteínas como nucleadoras e inhibidoras de HA, estas proteínas son ácidas y presentan altos porcentajes de estructura desordenada (Cross et al., 2000; Tye et al., 2003; Zurick et al., 2013) y la mayoría de las proteínas asociadas con la formación y el crecimiento de HA, son IDPs las cuales debido a la naturaleza flexible de sus estructuras pueden realizar diversas funciones: tienen la capacidad de estabilizar los iones o grupos de iones, el núcleo crítico, proporcionar sitios epitaxiales para la deposición inicial de minerales y/o estabilizar el cristal (Boskey and Villarreal-Ramírez, 2016).

Como han reportado Boskey y Villarreal-Ramírez en 2016, la permanencia del estado desordenado hace que la energía necesaria para unir el péptido a HA sea más favorable; no es una coincidencia que tantas proteínas involucradas en la biomineralización (OPN, BSP, DMP1, DPP y DSP) sean IDPs. La inestabilidad conformacional es una característica común de las proteínas que se unen a sólidos inorgánicos (Wojtas et al., 2012). El desorden intrínseco proporciona beneficios a estas proteínas y les permite cumplir sus funciones. Muchos IDPs involucrados en la biomineralización tienen una composición rica en residuos ácidos. Esta composición conduce a una fuerte repulsión electrostática y da como resultado la conformación extendida de la proteína, esta conformación proporciona una superficie de unión mucho más grande y es altamente conveniente, especialmente cuando interactúa con superficies cristalinas. Una composición extendida le permite a muchas proteínas actuar como inhibidores del crecimiento de los cristales al interactuar con la red cristalina y bloquear los sitios de nucleación, como es el caso de la proteína DPP (Wojtas et al., 2012). La conformación extendida combinada con el carácter ácido facilita las interacciones de las proteínas con los contraiones (Uversky, 2009). La unión a contraiones es crucial para debilitar la repulsión electrostática y permitir la formación de estructura secundaria una vez unida la proteína al mineral permitiendo las transiciones de desorden a orden (He et al., 2003).

Además una estructura desordenada confiere la capacidad de unirse a más de un ligando (otras proteínas asociadas al proceso de mineralización) y presentar diferentes superficies que facilitan la regulación de la mineralización, así mismo esta estructura abierta facilita modificaciones postraduccionales que pueden alterar la carga neta y proporcionar mayor variabilidad estructural, abriendo la puerta a otros ligando (Boskey and Villarreal-Ramírez, 2016; Uversky, 2009). Esta característica es impórtate puesto que las proteínas involucradas en la biomineralización tienen que unirse a muchos objetivos para cumplir adecuadamente sus funciones permitiendo el ensamblaje del complejo macromolecular crucial para la adecuada formación del biomineral.

Cabe mencionar que si bien se realizaron las simulaciones con diferentes péptidos estos no abarcaron la secuencia completa de CEMP1 y son de interés otras regiones que de acuerdo con nuestro análisis bioinformático cuentan con aminoácidos con carga positiva y negativa, posibles fosforilaciones, regiones con alto porcentaje de predicción de estructura secundaria y regiones con alto porcentaje de estructura desordenada donde se podrían encontrar otros motivos de unión a HA; así mismo en este estudio únicamente se abarcaron como modificaciones postraduccionales a las potenciales fosforilaciones y es de

relevancia para futuros estudios la consideración de otras modificaciones como las glicosilaciones, así como la formación de puentes disulfuro que estabilizan los elementos de estructura en las proteínas. De igual forma únicamente se abarco el eje cristalográfico (100) de la HA y pueden existir preferencias del motivo a superficies del cristal de HA, por lo que son importantes otros ejes cristalográficos como él (001) y (010). También dentro de consideraciones futuras se pudieran llevar acabo simulaciones con otras fases de la formación de HA, como podría ser con colágena o con fosfato octacálcico sin embargo para este último no existe actualmente un campo de fuerza adecuado.

Destacamos el péptido n°1 presenta los dos motivos de unión a HA, es de un tamaño pequeño y actualmente ya se ha utilizado para estudios *in vitro* como *in vivo* en la regeneración de tejido mineralizado(Correa et al., 2019). Además no existen estudios previos de SDM de la proteína CEMP1 en presencia de HA.

Conclusiones

- El mecanismo de interacción entre los péptidos derivados de CEMP1 y HA en el eje cristalográfico (100) es principalmente por interacción electrostática.

- Los dominios negativamente y positivamente cargados de los péptidos de CEMP1 son suficientes para lograr y dirigir la unión a la superficie del mineral de la hojuela de HA, pero la presencia de las fosforilaciones permite una unión más rápida y cercana a HA.

- Los motivos de unión cargados positivamente (HRR) y los motivos cargados negativamente generados por las fosforilaciones (TDS) funcionan a manera de "switch" para dirigir la unión a las caras de HA.

- Una vez que el los péptidos se unen a su ligando HA, los no fosforilados aumentan su estructura secundaria al unirse HA y solo si están fosforilados, permanecen en mayor porcentaje de estructura desordenada (arriba del 90%), lo cual le permitiría interactuar con otros ligando como son otras proteínas asociadas al proceso de mineralización.

- Las herramientas computacionales asi como de dinámica molecular ayudan a predecir una relación estructura-función de péptidos.

Bibliografía.

Ahmad, M., Hirz, M., Pichler, H., Schwab, H., 2014. Protein expression in Pichia pastoris: recent achievements and perspectives for heterologous protein production. Applied Microbiology and Biotechnology 98, 5301–5317. https://doi.org/10.1007/s00253-014-5732-5

Alberts Bruce, Johnson Alexander, Lewis Julian, Morgan David, Raff Martin, Roberts Keith, Walter Peter, 2016. Biología molecular de la célula, Sexta edición. ed. OMEGA, Barcelona.

Altschul, S.F., Boguski, M.S., Gish, W., Wootton, J.C., 1994. Issues in searching molecular sequence databases. Nature Genetics 6, 119–129. https://doi.org/10.1038/ng0294-119

AltschuP, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic Local Alignment Search Tool. Journal of Molecular Biology 215, 403–410.

Álvarez Pérez, M.A.A., Pitaru, S., Fregoso, O.A., Gasga, J.R., Arzate, H., 2003. Anti-cementoblastoma-derived protein antibody partially inhibits mineralization on a cementoblastic cell line. Journal of Structural Biology 143, 1–13. https://doi.org/10.1016/S1047-8477(03)00116-3

Álvarez-Pérez, M.A., Narayanan, S., Zeichner-David, M., Rodríguez Carmona, B., Arzate, H., 2006. Molecular cloning, expression and immunolocalization of a novel human cementum-derived protein (CP-23). Bone 38, 409–419. https://doi.org/10.1016/j.bone.2005.09.009

Anil K. Jain, Jianchang Mao, K. M. Mohiuddin, 1996. Artificial Neural Networks: A Tutorial. Comput. IEEE 31–44.

Arzate, H., Chimal-Monroy J, Hernández-Lagunas J, Diaz de León L, 1996. Human cementum protein extract promotes chondrogenesis and mineralization inmesenchymal cells. J Periodontal Res 31, 144–148.

Arzate, H., Jiménez-García, L.F., Álvarez-Pérez, M.A., Landa, A., Bar-Kana, I., Pitaru, S., 2002. Immunolocalization of a Human Cementoblastoma-conditioned Medium-derived Protein. Journal of Dental Research 81, 541–546. https://doi.org/10.1177/154405910208100808

Arzate, H., Zeichner-David, M., Mercado-Celis, G., 2015. Cementum proteins: role in cementogenesis, biomineralization, periodontium formation and regeneration. Periodontology 2000 67, 211–233. https://doi.org/10.1111/prd.12062

Basheer, I.A., Hajmeer, M., 2000. Artificial neural networks: fundamentals, computing, design, and application. Journal of Microbiological Methods 43, 3–31. https://doi.org/10.1016/S0167-7012(00)00201-3

Blom, N., Gammeltoft, S., Brunak, S., 1999. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. Journal of Molecular Biology 294, 1351–1362. https://doi.org/10.1006/jmbi.1999.3310

Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S., Brunak, S., 2004. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. PROTEOMICS 4, 1633–1649. https://doi.org/10.1002/pmic.200300771

Boskey, A.L., Villarreal-Ramírez, E., 2016. Intrinsically disordered proteins and biomineralization. Matrix Biology 52–54, 43–59. https://doi.org/10.1016/j.matbio.2016.01.007

Cai, Y.-D., Liu, X.-J., Chou, K.-C., 2003. Prediction of protein secondary structure content by artificial neural network. Journal of Computational Chemistry 24, 727–731. https://doi.org/10.1002/jcc.10222

Cai, Y.-D., Liu, X.-J., Xu, X., Chou, K.-C., 2002. Artificial neural network method for predicting protein secondary structure content. Computers & Chemistry 26, 347–350. https://doi.org/10.1016/S0097-8485(01)00125-5

Capriotti, L.A., Beebe, T.P., Schneider, J.P., 2007. Hydroxyapatite Surface-Induced Peptide Folding. Journal of the American Chemical Society 129, 5281–5287. https://doi.org/10.1021/ja070356b

Carmona-Rodríguez, B., Álvarez-Pérez, M.A., Narayanan, A.S., Zeichner-David, M., Reyes-Gasga, J., Molina-Guarneros, J., García-Hernández, A.L., Suárez-Franco, J.L., Chavarría, I.G., Villarreal-Ramírez, E., Arzate, H., 2007. Human Cementum Protein 1 induces expression of bone and cementum proteins by human gingival fibroblasts. Biochemical and Biophysical Research Communications 358, 763–769. https://doi.org/10.1016/j.bbrc.2007.04.204

Chou, Peter Y., Fasman, Gereald D., 1974. Prediction of Protein Conformation. Biochemistry 13, 222–245.

Chou, P.Y., Fasman, G.D., 1977. Secondary structural prediction of proteins from their amino acid sequence. Trends in Biochemical Sciences 2, 128–131. https://doi.org/10.1016/0968-0004(77)90440-6

Correa, R., Arenas, J., Montoya, G., Hoz, L., López, S., Salgado, F., Arroyo, R., Salmeron, N., Romo, E., Zeichner-David, M., Arzate, H., 2019. Synthetic cementum protein 1–derived peptide regulates mineralization *in vitro* and promotes bone regeneration *in vivo*. The FASEB Journal 33, 1167–1178. https://doi.org/10.1096/fj.201800434RR

Cregg, J.M., Cereghino, J.L., Shi, J., Higgins, D.R., 2000. Recombinant Protein Expression in Pichia pastoris. Molecular Biotechnology 16, 23–52. https://doi.org/10.1385/MB:16:1:23

Cross, K.J., Reynolds, E.C., Huq, N.L., 2000. Molecular Modelling of a Multiphosphorylated Sequence Motif Bound to Hydroxyapatite Surfaces. Journal of Molecular Modeling 6, 35–47. https://doi.org/10.1007/s0089400060035

Daly, R., Hearn, M., 2005. Expression of heterologous proteins inPichia pastoris: a useful experimental tool in protein engineering and production. Journal of Molecular Recognition 18, 119–138. https://doi.org/10.1002/jmr.687

Damasceno, L.M., Huang, C.-J., Batt, C.A., 2012. Protein secretion in Pichia pastoris and advances in protein production. Applied Microbiology and Biotechnology 93, 31–39. https://doi.org/10.1007/s00253-011-3654-z

David W. Mount, 2001. Bioinformatics: Sequence and Genome Analysis, 2da Edición. ed. CSHL Press.

Delano, W.L., 2002. The PyMOL Molecular Graphics System.

Dephoure, N., Gould, K.L., Gygi, S.P., Kellogg, D.R., 2013. Mapping and analysis of phosphorylation sites: a quick guide for cell biologists. Molecular Biology of the Cell 24, 535–542. https://doi.org/10.1091/mbc.e12-09-0677

Dingermann, T., 2008. Recombinant therapeutic proteins: Production platforms and challenges. Biotechnology Journal 3, 90–97. https://doi.org/10.1002/biot.200700214

Durrant, J.D., McCammon, J.A., 2011. HBonanza: A computer algorithm for molecular-dynamics-trajectory hydrogen-bond analysis. Journal of Molecular Graphics and Modelling 31, 5–9. https://doi.org/10.1016/j.jmgm.2011.07.008

Escobar, C.A.M., Murillo, L.V.R., Soto, J.F., 2011. Tecnologías bioinformáticas para el análisis de secuencias de ADN 6.

Ferron, F., Longhi, S., Canard, B., Karlin, D., 2006. A practical overview of protein disorder prediction methods. Proteins: Structure, Function, and Bioinformatics 65, 1–14. https://doi.org/10.1002/prot.21075

Foster, B.L., 2017. On the discovery of cementum. Journal of Periodontal Research 52, 666–685. https://doi.org/10.1111/jre.12444

Fujisawa, R., Mizuno, M., Nodasaka, Y., Yoshinori, K., 1997. Attachment of osteoblastic cells to hydroxyapatite crystals by a synthetic peptide (Glu7-Pro-Arg-Gly-Asp-Thr) containing two functional sequences of bone sialoprotein. Matrix Biology 16, 21–28. https://doi.org/10.1016/S0945-053X(97)90113-X

Garnier, Jean, Gibrat, Jean-Francois, Robson Barry, 1996. GOR Method for Predicting Protein Secondary Structure from Amino Acid Sequence. METHODS IN ENZYMOLOGY 266, 540–553.

George, A., Veis, A., 2008. Phosphorylated Proteins and Control over Apatite Nucleation, Crystal Growth, and Inhibition. Chemical Reviews 108, 4670–4693. https://doi.org/10.1021/cr0782729

Geourjon, C., Deléage, G., 1995. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Bioinformatics 11, 681–684. https://doi.org/10.1093/bioinformatics/11.6.681

Geourjon, C., Deléage, G., 1994. SOPM: a self-optimized method for protein secondary structure prediction. Protein Engineering, Design and Selection 7, 157–164. https://doi.org/10.1093/protein/7.2.157

Gilbert, M., Giachelli, C.M., Stayton, P.S., 2003. Biomimetic peptides that engage specific integrin-dependent signaling pathways and bind to calcium phosphate surfaces. Journal of Biomedical Materials Research 67A, 69–77. https://doi.org/10.1002/jbm.a.10053

Gilbert, M., Shaw, W.J., Long, J.R., Nelson, K., Drobny, G.P., Giachelli, C.M., Stayton, P.S., 2000. Chimeric Peptides of Statherin and Osteopontin That Bind Hydroxyapatite and Mediate Cell Adhesion. Journal of Biological Chemistry 275, 16213–16218. https://doi.org/10.1074/jbc.M001773200

Goldberg, H.A., Warner, K.J., Stillman, M.J., Hunter, G.K., 1996. Determination of the Hydroxyapatite-Nucleating Region of Bone Sialoprotein. Connective Tissue Research 35, 385–392. https://doi.org/10.3109/03008209609029216

Gomes, A.R., Byregowda, S.M., Veeregowda, B.M., Balamurugan, V., 2016. An Overview of Heterologous Expression Host Systems for the Production of Recombinant Proteins. Advances in Animal and Veterinary Sciences 4, 346–356. https://doi.org/10.14737/journal.aavs/2016/4.7.346.356

González, M.A., 2011. Force fields and molecular dynamics simulations. École thématique de la Société Française de la Neutronique 12, 169–200. https://doi.org/10.1051/sfn/201112009

Grzybowska, E., 2018. Calcium-Binding Proteins with Disordered Structure and Their Role in Secretion, Storage, and Cellular Signaling. Biomolecules 8, 42. https://doi.org/10.3390/biom8020042

Gustav-Stresemann, John von Neumann-Institut für Computing, Johannes Gutenberg-Universität Mainz, Max-Planck-Institut für Biophysikalische Chemie, Max-Planck-Institut für Polymerforschung (Eds.), 2004. Computational soft matter: from synthetic polymers to proteins, NIC series. NIC, Jülich.

He, G., Dahl, T., Veis, A., George, A., 2003. Dentin Matrix Protein 1 Initiates Hydroxyapatite Formation In Vitro. Connective Tissue Research 44, 240–245. https://doi.org/10.1080/03008200390181726

He, G., Ramachandran, A., Dahl, T., George, S., Schultz, D., Cookson, D., Veis, A., George, A., 2005. Phosphorylation of Phosphophoryn Is Crucial for Its Function as a Mediator of

Biomineralization. Journal of Biological Chemistry 280, 33109–33114. https://doi.org/10.1074/jbc.M500159200
Hollingsworth, S.A., Karplus, P.A., 2010. A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. BioMolecular Concepts 1. https://doi.org/10.1515/bmc.2010.022
Hughes, F.J., 2015. Periodontium and Periodontal Disease, in: Stem Cell Biology and Tissue Engineering in Dental Sciences. Elsevier, pp. 433–444. https://doi.org/10.1016/B978-0-12-397157-9.00038-2
Iakoucheva, L.M., 2004. The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Research 32, 1037–1049. https://doi.org/10.1093/nar/gkh253
Itoh, D., Yoneda, S., Kuroda, S., Kondo, H., Umezawa, A., Ohya, K., Ohyama, T., Kasugai, S., 2002. Enhancement of osteogenesis on hydroxyapatite surface coated with synthetic peptide (EEEEEEEPRGDT)in vitro. Journal of Biomedical Materials Research 62, 292–298. https://doi.org/10.1002/jbm.10338
Karplus, M., Petsko, G.A., 1990. Molecular dynamics simulations in biology. Nature 347, 631–639. https://doi.org/10.1038/347631a0
Komaki, M., Iwasaki, K., Arzate, H., Narayanan, A.S., Izumi, Y., Morita, I., 2012. Cementum protein 1 (CEMP1) induces a cementoblastic phenotype and reduces osteoblastic differentiation in periodontal ligament cells. Journal of Cellular Physiology 227, 649–657. https://doi.org/10.1002/jcp.22770
Kouza, M., Faraggi, E., Kolinski, A., Kloczkowski, A., 2017. The GOR Method of Protein Secondary Structure Prediction and Its Application as a Protein Aggregation Prediction Tool, in: Zhou, Y., Kloczkowski, A., Faraggi, E., Yang, Y. (Eds.), Prediction of Protein Secondary Structure. Springer New York, New York, NY, pp. 7–24. https://doi.org/10.1007/978-1-4939-6406-2_2
Kreegipuu, A., Blom, N., Brunak, S., 1999. PhosphoBase, a database of phosphorylation sites: release 2.0. Nucleic Acids Research 27, 237–239. https://doi.org/10.1093/nar/27.1.237
Krieger, E., Vriend, G., 2014. YASARA View—molecular graphics for all devices—from smartphones to workstations. Bioinformatics 30, 2981–2982. https://doi.org/10.1093/bioinformatics/btu426
Kumar, T.A., 2013. CFSSP: Chou and Fasman Secondary Structure Prediction Server. WIDE SPECTRUM: Research Journal. 1, 15–19.
Land, H., Humble, M.S., 2018. YASARA: A Tool to Obtain Structural Guidance in Biocatalytic Investigations, in: Bornscheuer, U.T., Höhne, M. (Eds.), Protein Engineering. Springer New York, New York, NY, pp. 43–67. https://doi.org/10.1007/978-1-4939-7366-8_4
Li, Jianzong, Feng, Y., Wang, X., Li, Jing, Liu, W., Rong, L., Bao, J., 2015. An Overview of Predictors for Intrinsically Disordered Proteins over 2010–2014. International Journal of Molecular Sciences 16, 23446–23462. https://doi.org/10.3390/ijms161023446
Li, Xiaohong, Romero, Pedro, Rani, Meeta, Dunker, A. Keith, Obradovic, Zoran, 1999. Predicting Protein Disorder for N-, C- and Internal Regions. Genome Informatics 10, 30–40.
Lindhe Jan, Lang P. Niklaus, Karring Thorkild, 2009. Periodontología Clínica e Implantología Odontológica, Quinta Edición. ed. Medica Panamericana, España.
Lobanov, M.Yu., Bogatyreva, N.S., Galzitskaya, O.V., 2008. Radius of gyration as an indicator of protein structure compactness. Molecular Biology 42, 623–628. https://doi.org/10.1134/S0026893308040195
López Villar, E., Nombela Cano, C., Røssel Larsen, M., 2008. Metodologías fosfoproteómicas útiles en estudios clínicos. Inmunología 27, 36–44. https://doi.org/10.1016/S0213-9626(08)70047-7

Lozano-Aponte, J., Scior, T., 2014. ¿Qué sabe Ud. Acerca de… Dinámica Molecular? Rev Mex Cienc Farm 3.

Macauley-Patrick, S., Fazenda, M.L., McNeil, B., Harvey, L.M., 2005. Heterologous protein production using thePichia pastoris expression system. Yeast 22, 249–270. https://doi.org/10.1002/yea.1208

Madden, T., 2002. The BLAST Sequence Analysis Tool.

Makrodimitris, K., Masica, D.L., Kim, E.T., Gray, J.J., 2007. Structure Prediction of Protein−Solid Surface Interactions Reveals a Molecular Recognition Motif of Statherin for Hydroxyapatite. Journal of the American Chemical Society 129, 13713–13722. https://doi.org/10.1021/ja074602v

Mark Abraham, Berk Hess, David van der Sopel, Erik Lindahl, 2014. GROMACS USER MANUAL Versión 5.0.4, Versión 5.0.4.

McGinnis, S., Madden, T.L., 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Research 32, W20–W25. https://doi.org/10.1093/nar/gkh435

Narayanan, K., Ramachandran, A., Hao, J., He, G., Park, K.W., Cho, M., George, A., 2003. Dual Functional Roles of Dentin Matrix Protein 1: IMPLICATIONS IN BIOMINERALIZATION AND GENE TRANSCRIPTION BY ACTIVATION OF INTRACELLULAR Ca $^{2+}$ STORE. Journal of Biological Chemistry 278, 17500–17508. https://doi.org/10.1074/jbc.M212700200

Newman Michael G., Carranza Fermin A., Takei Henry H., Klokkevold Perry R., 2018. Clinical Periodontology, Thirteenth Edition. ed. Elsevier, Saunders.

Paszun, D., Dominik, C., 2009. Collisional evolution of dust aggregates. From compaction to catastrophic destruction. Astronomy & Astrophysics 507, 1023–1040. https://doi.org/10.1051/0004-6361/200810682

Qin, C., Baba, O., Butler, W.T., 2004. Post-translational modifications of SIBLING proteins and their roles in osteogenesis and dentinogenesis. Critical Reviews in Oral Biology & Medicine 15, 126–136. https://doi.org/10.1177/154411130401500302

Rai, M., Padh, H., 2001. Expression systems for production of heterologous proteins. CURRENT SCIENCE 80, 8.

Ramachandran, G.N., Ramakrishnan, C., Sasisekharan, V., 1963. Stereochemistry of polypeptide chain configurations. Journal of Molecular Biology 7, 95–99. https://doi.org/10.1016/S0022-2836(63)80023-6

Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J., Dunker, A.K., 2001. Sequence complexity of disordered protein. Proteins: Structure, Function, and Genetics 42, 38–48. https://doi.org/10.1002/1097-0134(20010101)42:1<38::AID-PROT50>3.0.CO;2-3

Romo-Arévalo, E., Arzate, H., Montoya-Ayala, G., Rodríguez-Romero, A., 2016. High-level expression and characterization of a glycosylated human cementum protein 1 with lectin activity. FEBS Letters 590, 129–138. https://doi.org/10.1002/1873-3468.12032

Ruoslahti, E., 2003. The RGD story: a personal account. Matrix Biology 22, 459–465. https://doi.org/10.1016/S0945-053X(03)00083-0

Sawyer, A.A., Weeks, D.M., Kelpke, S.S., McCracken, M.S., Bellis, S.L., 2005. The effect of the addition of a polyglutamate motif to RGD on peptide tethering to hydroxyapatite and the promotion of mesenchymal stem cell adhesion. Biomaterials 26, 7046–7056. https://doi.org/10.1016/j.biomaterials.2005.05.006

Saygin, N.E., Giannobile, W.V., Somerman, M.J., 2000. Molecular and cell biology of cementum. Periodontology 2000 24, 73–98. https://doi.org/10.1034/j.1600-0757.2000.2240105.x

Serrano, J., Romo, E., Bermúdez, M., Narayanan, A.S., Zeichner-David, M., Santos, L., Arzate, H., 2013. Bone Regeneration in Rat Cranium Critical-Size Defects Induced by Cementum Protein 1 (CEMP1). PLoS ONE 8, e78807. https://doi.org/10.1371/journal.pone.0078807

Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. Journal of Molecular Biology 147, 195–197. https://doi.org/10.1016/0022-2836(81)90087-5

Tompa, P., 2012. Intrinsically disordered proteins: a 10-year recap. Trends in Biochemical Sciences 37, 509–516. https://doi.org/10.1016/j.tibs.2012.08.004

Tye, C.E., Rattray, K.R., Warner, K.J., Gordon, J.A.R., Sodek, J., Hunter, G.K., Goldberg, H.A., 2003. Delineation of the Hydroxyapatite-nucleating Domains of Bone Sialoprotein. Journal of Biological Chemistry 278, 7949–7955. https://doi.org/10.1074/jbc.M211915200

Uversky, V.N., 2009. Intrinsically Disordered Proteins and Their Environment: Effects of Strong Denaturants, Temperature, pH, Counter Ions, Membranes, Binding Partners, Osmolytes, and Macromolecular Crowding. The Protein Journal 28, 305–325. https://doi.org/10.1007/s10930-009-9201-4

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E., Berendsen, H.J.C., 2005. GROMACS: Fast, flexible, and free. Journal of Computational Chemistry 26, 1701–1718. https://doi.org/10.1002/jcc.20291

Villarreal-Ramírez, E., Eliezer, D., Garduño-Juarez, R., Gericke, A., Perez-Aguilar, J.M., Boskey, A., 2017. Phosphorylation regulates the secondary structure and function of dentin phosphoprotein peptides. Bone 95, 65–75. https://doi.org/10.1016/j.bone.2016.10.028

Villarreal-Ramírez, E., Moreno, A., Mas-Oliva, J., Chávez-Pacheco, J.L., Narayanan, A.S., Gil-Chavarría, I., Zeichner-David, M., Arzate, H., 2009. Characterization of recombinant human cementum protein 1 (hrCEMP1): Primary role in biomineralization. Biochemical and Biophysical Research Communications 384, 49–54. https://doi.org/10.1016/j.bbrc.2009.04.072

Wojtas, M., Dobryszycki, P., Oyhar, A., 2012. Intrinsically Disordered Proteins in Biomineralization, in: Seto, J. (Ed.), Advanced Topics in Biomineralization. InTech. https://doi.org/10.5772/31121

Wright, P.E., Dyson, H.J., 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. Journal of Molecular Biology 293, 321–331. https://doi.org/10.1006/jmbi.1999.3110

Xue, B., Dunbrack, R.L., Williams, R.W., Dunker, A.K., Uversky, V.N., 2010. PONDR-FIT: A meta-predictor of intrinsically disordered amino acids. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics 1804, 996–1010. https://doi.org/10.1016/j.bbapap.2010.01.011

Yin, J., Li, G., Ren, X., Herrler, G., 2007. Select what you need: A comparative evaluation of the advantages and limitations of frequently used expression systems for foreign genes. Journal of Biotechnology 127, 335–347. https://doi.org/10.1016/j.jbiotec.2006.07.012

Zurick, K.M., Qin, C., Bernards, M.T., 2013. Mineralization induction effects of osteopontin, bone sialoprotein, and dentin phosphoprotein on a biomimetic collagen substrate. Journal of Biomedical Materials Research Part A 101A, 1571–1581. https://doi.org/10.1002/jbm.a.34462