



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE QUÍMICA

IDENTIFICACIÓN DE DOMINIOS DE UNIÓN A CALCIO Y EL EFECTO
DE LAS FOSFORILACIONES EN LA SIALOPROTEÍNA ÓSEA,
MEDIANTE ANÁLISIS BIOINFORMÁTICOS Y SIMULACIONES DE
DINÁMICA MOLECULAR

TESIS

QUE PARA OBTENER EL TÍTULO DE
QUÍMICO FARMACÉUTICO BIÓLOGO

PRESENTA

RODOLFO MARTÍNEZ CARRANZA

CD. MX.

2019





Universidad Nacional
Autónoma de México



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

JURADO ASIGNADO:

PRESIDENTE: Profesor: Jesús Fernando Montiel Aguirre

VOCAL: Profesor: José Luis Medina Franco

SECRETARIO: Profesor: Eduardo Villarreal Ramírez

1er. SUPLENTE: Profesor: Rodrigo Aguayo Ortiz

2° SUPLENTE: Profesor: Audifás Salvador Matus Meza

SITIO DONDE SE DESARROLLÓ EL TEMA: LABORATORIO DE BIOINGENIERÍA DE TEJIDOS, DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN (DEPEI) DE LA FACULTAD DE ODONTOLOGÍA, UNAM

ASESOR DEL TEMA: Dr. Eduardo Villarreal Ramírez _____

SUSTENTANTE: Rodolfo Martínez Carranza _____

Agradecimientos

Se agradece al Dr. Eduardo Villarreal Ramírez, por aceptarme como su alumno y permitir la realización de este trabajo, así como el servicio social.

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

Rodolfo Martínez Carranza agradece al proyecto PAPIIT/UNAM: IA207218 por la beca otorgada.

Se agradece al Dr. Luis Lozano Aguirre Beltrán por la escritura de códigos en lenguaje PERL para el análisis de secuencias.

Índice

1.0	Introducción.....	6
2.0	Objetivos.....	11
2.1	Objetivo general.....	11
2.2	Objetivos particulares.....	11
3.0	Antecedentes.....	12
3.1	Biom mineralización.....	12
3.1.1	Biom mineralización biológicamente inducida.....	12
3.1.2	Biom mineralización biológicamente controlada.....	13
3.1.3	Tejidos mineralizados humanos.....	15
3.2	IDP.....	18
3.2.1	SIBLINGs.....	20
3.2.2	BSP.....	20
4.0	Metodología y bases teóricas.....	24
4.1.1	BLAST.....	24
4.1.2	Aprendizaje automatizado (ML).....	29
4.1.3	Redes neurales artificiales (NN).....	30
4.1.4	Máquinas de vectores de soporte (SVM).....	33
4.1.5	Predicción de fosforilaciones.....	37
4.1.6	Predicción de IDR.....	38
4.1.7	Predicción de estructura secundaria.....	41
4.1.8	Alineamiento múltiple de secuencias (MSA).....	45
4.2	Simulaciones de DM.....	48
4.2.1	Integrador <i>Leap-frog</i>	49
4.2.2	Campo de fuerza.....	50
4.2.3	Minimización de energía.....	52
4.2.4	Ensamble NVT.....	54
4.2.5	Ensamble NPT.....	55
4.2.6	GROMACS.....	56
5.0	Resultados.....	57
5.1	Bioinformática.....	57
5.1.1	Predicción de fosforilaciones.....	57

5.1.2 Predicción de IDR's	57
5.1.3 Predicción de estructura secundaria.....	58
5.1.4 Alineamiento múltiple de secuencias (MSA)	60
5.1.5 Selección de péptidos	63
5.2 Simulaciones de DM.....	65
5.2.1 Péptido 1	66
5.2.2 Péptido 2	75
5.2.3 Péptido 3	79
5.2.4 Péptido 4	87
5.2.5 Péptido 5	94
5.2.6 Péptido 6	100
6.0 Discusión.....	109
6.1 Selección de péptidos.....	109
6.2 Simulaciones de DM.....	112
7.0 Conclusiones	118
8.0 Perspectivas.....	119
9.0 Referencias	120
10.0 Anexos	134
Anexo 1: Nomenclatura de aminoácidos	134
Anexo 2: Secuencias de BSP	135
Anexo 3: Validación, sensibilidad y especificidad de servidores	141

1.0 Introducción

La biomineralización se define como la formación de depósitos minerales (estructuras cristalinas) en los seres vivos. La formación de los minerales comienza a escalas atómicas llevando a la formación de bloques nanoestructurados macroscópicos y organizados en estructuras jerárquicas complejas (Beniash, 2011). Los minerales formados en el proceso de biomineralización, también conocidos como “biominerales”, tienen distintas funciones en los organismos como: protección, locomoción, mecanotransducción, masticación, reservorio de iones y funciones ópticas. Por ejemplo, las conchas de moluscos o los huesos de la caja torácica tienen una función de protección, mientras los órganos dentarios participan en funciones masticatorias y huesos largos tienen funciones de soporte, por citar algunos (Chen, 2019).

La biomineralización es un área de estudio multidisciplinaria que involucra a la biología, química, cristalografía, ciencia de materiales, mineralogía, entre otras ramas científicas (Estroff, 2008). La importancia radica en sus potenciales aplicaciones a largo plazo, una vez entendidos los mecanismos biológicos asociados a la formación de los biominerales. Dichas aplicaciones abarcan los campos de la bioingeniería de tejidos en la síntesis de biomateriales implantables o incluso la ingeniería ecológica en el diseño de materiales sustentables (Ige, 2012). Además, con vital importancia en el tratamiento de patologías asociadas a los tejidos mineralizados, como la osteoporosis u osteogénesis imperfecta (Sun, 2019).

Los mamíferos presentan diferentes tejidos mineralizados, como el tejido óseo, esmalte, dentina, cemento radicular y otolitos (Pasteris, 2008). El hueso es el tejido mineralizado más abundante en el cuerpo humano, cuya composición química y organización estructural son responsables de sus propiedades mecánicas únicas. El hueso es un material compuesto heterogéneo, formado por una fase mineral y una fase orgánica, esta última también conocida como matriz orgánica. La fase mineral está formada por el mineral de hidroxiapatita $[\text{Ca}_{10}(\text{PO}_4)_6(\text{OH})_2]$ (HA) y la

fase orgánica se divide en proteínas colágenas (~90%), proteínas no colágenas (NCP)(~5%) lípidos (~2%) y agua (Boskey, 2013).

La mayoría de las proteínas que conforman las NCP son proteínas intrínsecamente desordenadas (IDP) (Wojtas, 2012). Las IDP se caracterizan por contener en su secuencia de aminoácidos un alto contenido de residuos con carga eléctrica y polares, pocos residuos aromáticos e hidrofóbicos, ser altamente flexibles, poder unirse a más de un ligando y tener múltiples funciones (Dyson, 2016). Las IDP no tienen una estructura tridimensional definida y adquieren una conformación de giros aleatorios y bucles (*loops*), ya sea en la totalidad de la cadena peptídica o en regiones localizadas; estas regiones reciben el nombre de intrínsecamente desordenadas (IDR). Experimentalmente, es complejo dilucidar la estructura de una IDP, prueba de ello es la existencia de proteínas en la *Protein Data Bank* (PDB) con regiones faltantes debido a su muy baja resolución y la alta movilidad del segmento (Uversky, 2010). Las IDP carecen de la capacidad de formar cristales de proteína y no dispersan de forma adecuada los rayos X (Uversky, 2010). Además, las IDP relacionadas a la formación de los tejidos mineralizados son de un tamaño mayor a los 35 kDa, lo cual es el límite en resolución de la resonancia magnética nuclear (NMR) de proteínas (Frueh, 2013). Una aproximación para obtener un panorama atómico de las IDP, son los métodos de simulaciones de dinámica molecular (DM) para estudiar la estructura tridimensional de las IDP y las interacciones con sus ligandos (Perovic, 2018). Sin embargo, las IDP pueden adquirir estructuras estables al ser modificadas postraduccionalmente o al unirse a alguno de sus ligandos (Boskey, 2007).

Las IDP participan en procesos de señalización celular, regulación de la transcripción, traducción y acoplamiento de ciclo celular, organización de la cromatina, transporte nuclear y en procesos patológicos como la metástasis del cáncer (Wright, 2015; Kalmar, 2012).

Las NCP asociadas al tejido óseo y órganos dentarios tienen un mayor número de fosforilaciones, a diferencia de otras NCP de tejidos conectivos suaves (Veis, 1997).

Esto sugiere que las fosforilaciones son parte importante del mecanismo de biomineralización.

Inicialmente, las NCP encontradas en la matriz orgánica del hueso fueron estudiadas por su capacidad de inducir la adhesión celular, unión a iones Ca^{2+} e hidroxiapatita, y participar en la formación del tejido óseo (Young, 1992). Actualmente, se sabe que las NCP de los tejidos mineralizados consisten principalmente en proteínas ácidas y tienen un rol crucial en el proceso de biomineralización, dada su capacidad de regular al proceso positiva o negativamente (Gorski, 2011). Las NCP actúan como nucleadoras, reguladoras e inhibidoras de la orientación y crecimiento de los cristales minerales de HA (Kalka, 2019).

Una de las NCP de mayor interés es la sialoproteína ósea (BSP), la cual representa aproximadamente un 12% de las NCP en tejidos mineralizados (McKee, 2012). La BSP fue reportada por primera vez en 1972, sin embargo, sólo se trataba de un fragmento de la proteína. La BSP completa se purificó a partir de hueso bovino, se describió como una proteína con una masa molecular relativa (Mr) de 59,000, un alto contenido de ácido siálico y rica en oligosacáridos (Heinegård, 1989). En general, BSP es una glicoproteína de ~68 kDa (Kim, 1994) codificada por el gen IBSP localizado en el cromosoma 4 en humanos (Fisher, 1990). Su estructura primaria consta de trescientos diecisiete aminoácidos (Anexo 2), de los cuales los dieciséis primeros aminoácidos corresponden a la péptido señal (Ganss, 1999) y se le han reportado distintas funciones. BSP contiene una secuencia de reconocimiento de integrinas, un dominio RGD (arginina-glicina-aspartato), el cual es crítico para la señalización celular y en la diferenciación de los osteoblastos (Baht, 2008). Se ha mostrado que BSP puede unirse a distintos ligandos como colágeno, metaloproteinasas, integrinas e hidroxiapatita (Gordon, 2007). Además, mediante experimentos realizados con $^{45}\text{Ca}^{2+}$ se demostró la capacidad de interacción de BSP con iones Ca^{2+} , cuyos resultados indican una unión a ochenta y tres iones Ca^{2+} por molécula y una constante de unión de 0.5-0.1mM (Chen, 1992). Mediante una modificación del método de Scatchard y usando BSP, purificada de

una línea celular de osteosarcoma de rata, se calculó una $K_d \approx 2.6 \times 10^{-9}M$ para la interacción BSP con HA (Stubbs, 1997). Adicionalmente, se demostró el potencial de BSP con modificaciones postraduccionales para inhibir la nucleación de HA (Stubbs, 1997). A pesar de estos datos, Hunter (1993) demostró que BSP tiene la facultad de nuclear cristales de HA, utilizando BSP purificada de la calvaria de cerdos adultos. Las IDP de los tejidos mineralizados, pueden actuar de manera distinta en la nucleación de los minerales debido a la posición de las modificaciones postraduccionales, la concentración de la proteína y su origen (Villarreal-Ramirez, 2017).

BSP en mamíferos tiene una conservación en secuencia del 45% y entre un 10% y 23% de sustituciones conservativas (Ganss, 1999). Las regiones con mayor conservación en la secuencia corresponden a los extremos amino y carboxilo terminal, la secuencia RGD y las regiones ricas en ácido glutámico. Es generalmente aceptado que la conservación filogenética de una secuencia suele implicar su importancia biológica funcional y estructural (Ponting, 2017). Estudios de la estructura de BSP usando NMR unidimensional concluyeron que BSP tiene una estructura flexible sin regiones estructurales significantes (Fisher, 2001). Los resultados de dicroísmo circular (CD) reportados por Wuttke (2001) indican un contenido de 5% hélice α , 32% lámina β , 17% giros β y 46% giros aleatorios. Mientras, los estudios reportados por Tye (2003) mostraron en CD para BSP recombinante un contenido en estructura secundaria de 13.8% de hélice α , 43.5% de lámina β , 10.8% de giros β , 8.1% de hélice poli-pro (II) y 18.9% de estructura desordenada. Además, en estudios de dispersión de rayos X de bajo ángulo (SAXS) se observó una estructura de giros aleatorios (Tye, 2003).

BSP es una proteína con una carga global negativa y con un alto grado de fosforilación. Experimentalmente, se ha demostrado que BSP tiene, de manera natural (*in vivo*), aproximadamente 5.8 mol de serina fosforilada por mol de proteína, modificaciones correspondientes a las enzimas CKII, PKC y cinasa dependiente de cGMP (Salih, 2003). Estudios realizados en BSP bovina muestran que hay 11 sitios potenciales identificados para ser fosforilados, sin embargo, no todos se encuentran

fosforilados para una población dada de moléculas aisladas de hueso a una edad específica, mostrando así una heterogeneidad en las modificaciones postraduccionales (George, 2008). Además, se sabe que la mayoría de estas fosforilaciones ocurren en la región cercana al extremo amino terminal. Este cluster de fosforilaciones puede influenciar la capacidad de unión a HA de dos zonas de repeticiones de ácido glutámico debido a la alta carga negativa (Ganss, 1999).

Debido a lo anterior esta proteína se señala como un posible punto de control en el proceso de formación de tejido mineralizado. Sin embargo, la naturaleza de sus interacciones y mecanismos no son completamente conocidos debido a la falta de una estructura experimental a nivel atómico (Yang, 2010).

No obstante, existen estudios de simulaciones de DM que permiten ampliar el conocimiento acerca de las posibles interacciones de la BSP. Tal es el caso de trabajo realizado por Vincent (2013), donde un modelo de BSP fue construido con base en la información existente en publicaciones incluyendo glicosilaciones, fosforilaciones y sulfataciones. Con los resultados de este trabajo se sugiere que las moléculas de BSP son transferidas de la superficie celular a las fibras de colágeno donde actúan como nucleador de HA por medio de interacciones electrostáticas con iones Ca^{2+} en la fase mineral. Sin embargo, no profundizan en las interacciones específicas entre los iones Ca^{2+} y BSP. En la investigación de Baht (2010), se estudia el rol la serina 136 fosforilada en la BSP de rata (serina 133 en BSP humana) sobre la nucleación de HA y se encontró que la fosforilación regula la nucleación de HA.

En esta tesis se busca estudiar distintos fragmentos de la sialoproteína ósea humana para evaluar el efecto de las fosforilaciones sobre su capacidad de interacción con Ca^{2+} y sus propiedades estructurales para la identificación de dominios que puedan estar involucrados en el proceso de biomineralización.

2.0 Objetivos

2.1 Objetivo general

Identificar los motivos de unión a Ca^{2+} y el efecto de las fosforilaciones en los péptidos de la sialoproteína osea.

2.2 Objetivos particulares

- 1) Realizar mediante análisis bioinformático una selección de péptidos correspondientes a regiones de la BSP cuyas características pudieran permitir la interacción con sus ligandos fisiológicos.
- 2) Determinar los dominios de unión a Ca^{2+} en la proteína BSP mediante un método de simulaciones de dinámica molecular.
- 3) Correlacionar los cambios estructurales de los péptidos de BSP en ausencia y presencia de su ligando Ca^{2+} .
- 4) Identificar el efecto de las fosforilaciones en la BSP y en la asociación a su ligando Ca^{2+} .

3.0 Antecedentes

3.1 Biomineralización

La biomineralización es un fenómeno ampliamente distribuido en la naturaleza, cuyo resultado es la formación de minerales a partir de materia orgánica e inorgánica. Estos minerales tienen diversas funciones en los organismos donde fueron producidos: brindar protección como las conchas de los moluscos y braquiópodos, funciones masticatorias y dar soporte como los dientes y huesos en mamíferos (Addadi, 2014), por mencionar algunos ejemplos. Actualmente la biomineralización es todo un campo de estudio multidisciplinario relacionado con ciencias como medicina, biogeoquímica, bioquímica, ingeniería de tejidos y biomateriales.

La biomineralización puede ser clasificada de acuerdo con varios criterios, uno de ellos es el grado de control que ejerce un organismo sobre ésta. Resultando en “Biomineralización biológicamente inducida” y “Biomineralización biológicamente controlada” (Lowenstam, 1981).

3.1.1 Biomineralización biológicamente inducida

La formación de minerales por este tipo de proceso generalmente tiene como característica una nucleación y crecimiento extracelular, siendo resultado de la actividad metabólica del organismo, así como interacciones químicas entre subproductos y el medio ambiente, carencia de morfología y orientación cristalina, y carencia de algún mecanismo regulador específico (Lowenstam, 1989).

Este fenómeno es predominante en bacterias y hongos, aunque esto no significa que sea inexistente en animales o plantas. Un ejemplo es la mineralización de la corteza del árbol *Milicia excelsa*, un árbol originario de África, como consecuencia de la oxidación de oxalato de calcio, debida a bacterias, y su posterior deposición en tejidos (Braissant, 2004).

3.1.2 Biomineralización biológicamente controlada

La producción de minerales como un proceso controlado es consecuencia de actividades fisiológicas del organismo y se encuentra controlada en tres aspectos claves (Chen, 2019):

- Espacial: Ocurre en lugares delimitados como pueden ser vesículas o uniones estrechas entre células.
- Estructural: Los cristales presentan un tamaño y forma uniformes, y un arreglo cristalino definido, a diferencia de los cristales formados por un proceso inducido.
- Químico: La composición del cristal, y por lo tanto sus propiedades, es constante. El mineral no se ve afectado por moléculas circundantes a diferencia de la mineralización inducida, donde el medio ambiente puede influir en la formación de uno u otro compuesto.

El proceso puede ser descrito de diferentes formas, de acuerdo con los autores y a la postura o finalidad con la que se aborda el tema, este trabajo se apega a la descripción realizada por Lowenstam (1989). En este libro se menciona que un proceso de mineralización biológicamente controlada puede ser descrito en función de los distintos procesos de control que tienen lugar para obtener un producto final. Teniendo en consideración que “en el mundo real” las etapas pueden ser muy diferentes o incluso simultáneas en función del organismo. A continuación, se habla brevemente de estos puntos de control:

- Restricción espacial.

El lugar en el que se forma el mineral es aislado del resto del ambiente, ya sea por vesículas, bicapas lipídicas, células o polímeros de macromoléculas. Dichas restricciones pueden ser intra y extracelulares. Un ejemplo claro de restricción espacial se encuentra en las bacterias que producen magnetosomas, también conocidas como bacterias magnetotácticas. Un magnetosoma es una vesícula

intracelular en la que un cristal de un mineral con propiedades magnéticas presenta nucleación y crecimiento (Gorby, 1988).

- Matriz orgánica preformada.

Principalmente en aquellos procesos involucrados con el desarrollo de partes esqueléticas se observa una subdivisión del espacio previamente definido, regularmente por células, para la mineralización. Esta subdivisión corresponde a un conjunto de macromoléculas conocidas como la matriz orgánica. Es necesario aclarar que no todos los casos en los que hay una restricción espacial se presenta una matriz orgánica, pero para los propósitos de este trabajo es el escenario ideal.

La matriz orgánica se forma por el autoensamblaje de moléculas secretadas al espacio previamente delimitado. Un ejemplo es la matriz orgánica compuesta por moléculas de colágeno y otras proteínas en la formación de hueso y dientes (Miller, 1984). Las matrices tienen la característica de ser permeables a los iones y ayudar a formar la estructura en la que se llevará a cabo la mineralización.

- Saturación del medio.

Uno de los requisitos indispensables para la precipitación del mineral es la saturación del ambiente en el que se llevará a cabo, así como todas las consideraciones cinéticas y termodinámicas necesarias. Estas condiciones son idénticas a aquellas involucradas en el estudio de soluciones saturadas en el laboratorio (Nancollas, 1977). La formación del licor madre se encuentra mediada por toda la maquinaria celular encargada de regular la biomineralización. Ejemplo de esto son las bombas de iones específicas.

- Nucleación.

El mecanismo de nucleación puede variar dependiendo de las características del proceso de mineralización, específicamente del tipo de delimitación. A manera general, puede existir un control indirecto o directo sobre ésta. De manera indirecta, la presencia de ciertas moléculas en solución puede inhibir la formación de un núcleo naciente de un mineral y, a la vez, permitir que otro se forme. Por ejemplo,

la presencia de ATP inhibe la formación de fosfato de calcio cristalino y favorece la formación de fosfato de calcio amorfo (Posner, 1984). Mientras que el control directo suele ocurrir en una superficie sólida pues facilita la formación del núcleo crítico al limitar el movimiento de las moléculas; sin embargo, se carece de detalles a nivel nanométrico de las interacciones existentes en etapas tempranas (nucleación y crecimiento) por la falta de herramientas que permitan generar imágenes de alta resolución en ambientes fluidos (Kröger, 2015).

- Crecimiento y terminación del cristal.

Hay estudios *in vivo* realizados en erizo de mar donde se muestra que la regulación del crecimiento del cristal esta mediada por glicoproteínas ácidas que a la vez afectan la morfología del cristal que se forma (Lowenstam, 1989). El mineral crece sobre las proteínas y se observa que estas se encuentran ocluidas dentro del cristal ya formado. Existen pocos ejemplos detallados del proceso de crecimiento de un cristal; pero hay teorías que indican a las moléculas encargadas de favorecer la nucleación del cristal como responsables del proceso terminación de éste. Un ejemplo son los polisacáridos involucrados en la formación de cocolitos. Estos polisacáridos se encuentran ligados a la membrana de vesículas donde interactúan con proteínas e inducen el crecimiento de los cristales, posteriormente al presentarse como moléculas libres cubren al cocolito e inhiben el crecimiento terminando así la formación del mineral (Simkiss, 2012).

3.1.3 Tejidos mineralizados humanos

Los tejidos mineralizados son estructuras compuestas de una fase que comprende a la matriz orgánica, una fase mineral y células cuyas características, como la complejidad y composición, pueden variar de acuerdo con el tipo de tejido y el organismo (Kruzic, 2008). Algunos ejemplos son huesos, dientes, esponjas de mar, nácar y conchas de moluscos. Para el propósito de este trabajo sólo se abordarán aquellos presentes en humanos haciendo especial énfasis en huesos. Sin embargo, se pretende dejar de lado detalles anatómicos/fisiológicos que no sean plenamente necesarios para el cometido de esta tesis.

Boskey (2007) realiza un análisis jerárquico de la estructura de huesos y dientes, desde la escala de centímetros (órganos) hasta la ultraestructura (componentes individuales de los tejidos). Con esta información, junto a la publicación de McKee (2005), se pretende dar un contexto que sirva como prefacio para las siguientes explicaciones.

- Escala de órganos.

Es anatómicamente posible diferenciar entre distintos tipos de huesos, mismos que tienen diferentes funciones y estructuras variables. Por ejemplo, se puede distinguir entre los componentes del diente: dentina, cemento y esmalte.

- Escala tisular.

A este nivel se distinguen tres componentes fundamentales: una fase mineral, una matriz orgánica y células que controlan el proceso de mineralización (sección 3.2.2). Una de las funciones de estas células es mantener los niveles de iones calcio, magnesio y fosfato en el cuerpo, constituyendo de esta manera una de las mayores funciones no mecánicas de los huesos. Cuando hay un desequilibrio en esta homeostasis puede ser señal de alguna patología.

- Escala microestructural.

En esta escala se destacan las unidades estructurales que componen a los tejidos, como trabéculas y osteonas en hueso y túbulos que permean la dentina en los dientes.

- Escala ultraestructural.

En este nivel se encuentran los elementos de mayor interés para este trabajo. Esta escala engloba los componentes individuales de los tejidos: matriz orgánica y los cristales minerales. El biomineral que se encuentra en dientes y huesos es un análogo al mineral geológico HA.

La HA geológica o HA estequiométrica es un mineral de la familia de las apatitas, su fórmula ideal es $\text{Ca}_{10}(\text{PO}_4)_6(\text{OH})_2$ con relación molar de 1.67 Ca/P y es una de

las formas más usuales del fosfato de calcio (Fihri, 2017). Desde un punto de vista químico las principales diferencias entre la HA biológica, o no estequiométrica, y la estequiométrica son: un alto grado de sustitución en las redes cristalinas donde cationes como Na^+ , Mg^{2+} , K^+ , entre otros, toman el lugar de iones Ca^{2+} , la presencia de carbonatos, deficiencia en grupos OH^- y diferencias de solubilidad (Lin, 2015). Los cristales de HA son depositados en la matriz orgánica, que funciona a modo de andamio para la formación de estos. La matriz orgánica se encuentra constituida principalmente de colágeno.

El colágeno es una familia de proteínas estructurales que se encuentra ampliamente distribuida en prácticamente todos los eucariontes, a excepción de las plantas y protozoarios, y que en el humano representa la proteína más abundante constituyendo más de un cuarto de la masa seca del cuerpo (Collier, 2018). Una molécula de colágeno está formada por tres cadenas polipeptídicas conocidas como cadenas α . De los distintos tipos de colágeno descritos el tipo I es el más abundante, encontrándose en diversos tejidos como tendones, ligamentos, hueso, piel y corneas. Es un heterotrímero de composición $\{[\alpha_1(\text{I})]_2 [\alpha_2(\text{I})]\}$, es decir dos cadenas α_1 y una cadena α_2 de colágeno tipo I, que son codificadas por los genes COL1A1 y COL1A2, respectivamente (Shoulders, 2009). Cada una de las cadenas tiene repeticiones Gly-X-Y donde X y Y pueden ser cualquier aminoácido, pero con mayor frecuencia son prolina y 4-hidroxiprolina, respectivamente. La presencia de estos aminoácidos es responsable de la característica triple hélice en las moléculas de colágeno, siendo la 4-hidroxiprolina esencial para la formación de puente de hidrógeno intramoleculares que contribuyen a la estabilidad de este dominio helicoidal (Kadler, 1996). Las moléculas de colágeno también cuentan con dominios en los extremos N y C terminales que no forman triple hélice y son conocidos como “dominios no colágenos (NC)”, los cuales se numeran, en función del carboxilo terminal, como NC1, NC2, etc (Kadler, 2007).

El colágeno tipo I se encuentra clasificado como “formador de fibrillas”, esto significa que moléculas helicoidales individuales, conocidas como tropocolágeno, se ensamblan en un complejo jerárquico. El rasgo característico de las fibrillas es que

tienen una periodicidad “ D ”, donde $D = 67\text{nm}$. Este término hace referencia al escalonamiento axial de las moléculas individuales, esta distancia es la suma de las regiones de separación (*gaps*) y de superposición (Shoulders, 2009), tal como se muestra en la figura 1.

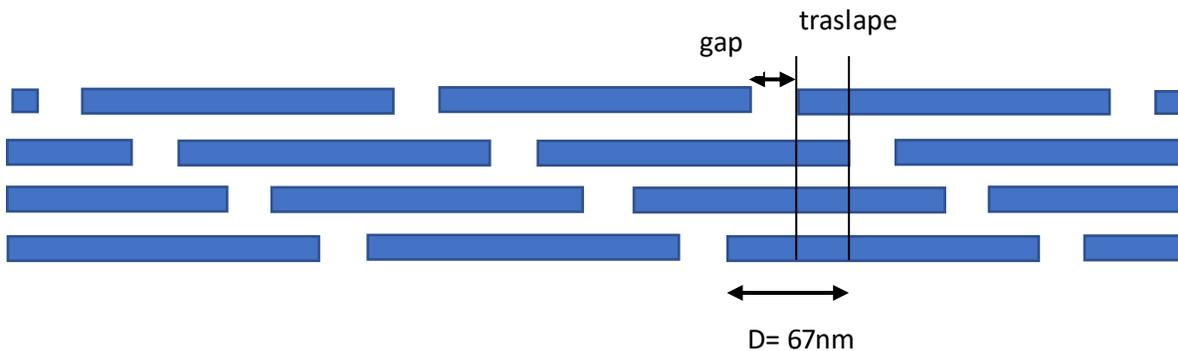


Figura 1. Representación esquemática de una fibrilla de Colágeno

El arreglo de las moléculas de tropocolágeno se representa de color azul, se ejemplifica un gap y una zona de traslape. El gap equivale a $0.54D$ y el traslape a $0.46D$.

Este arreglo es responsable de la flexibilidad que presentan los tejidos no mineralizados, mientras que al ser cubierto de mineral brinda la fuerza característica de un tejido mineralizado. Se sabe que las partículas de HA se depositan en estos *gaps* para después expandirse a lo largo de la fibrilla y de la matriz orgánica.

Sin embargo, el colágeno no es la única molécula que conforma la matriz orgánica. También está compuesta por proteínas “no colágenas” (NCP). Existen estudios sobre estas proteínas donde se señala su capacidad para unirse al colágeno y a la HA, indicando un rol potencial en el desarrollo de hueso durante la mineralización (Termine, 1981). Estas NCP frecuentemente son clasificadas como proteínas intrínsecamente desordenadas (IDP).

3.2 IDP

Las proteínas desempeñan un amplio espectro de funciones: enzimas, hormonas, antibióticos, acarreadores de electrones, anticuerpos, por mencionar algunas (McLachlan, 1972). Se sabe que la estructura de una proteína está determinada por su secuencia de aminoácidos y comúnmente, en la enseñanza, se menciona

que una proteína puede adoptar estructuras de tipo secundaria, terciaria y en algunos casos estructura cuaternaria. También se sabe que un dominio proteico es una subestructura que se forma de partes de una cadena peptídica que se puede plegarse en una estructura estable y compacta (Alberts, 2015). Esto es un paradigma que surgió en el siglo XX gracias a la solidificación del conocimiento sobre la relación función-estructura de las proteínas, donde el punto central señala que una estructura tridimensional es un requisito obligatorio para la función de una proteína (Romero, 2001). Sin embargo, en el año de 1978 mientras se estudiaba la proteína disco del virus mosaico del tabaco se descubrió una región descrita como “un *loop* flexible” de aproximadamente veinticuatro aminoácidos, la cual a pesar de carecer de una estructura definida tenía la capacidad de interactuar con RNA (Bloomer, 1978). Con el paso de los años se incrementó el estudio de estas proteínas, conocidas como IDPs, y desde la década de 1990 se han vuelto un área de especial interés debido a que cada vez se conocen más papeles funcionales de estas moléculas tanto en procesos biológicos como en patológicos, y hoy en día se sabe que representan aproximadamente un 30% del proteoma humano (Atkins, 2015). Boskey (2016) señala que las IDP son proteínas con una estructura altamente variable, algunas de ellas no presentan una estructura definida mientras que otras presentan regiones desordenadas. Exhiben un alto grado de hidrofiliidad debido a su estructura primaria, siendo rica en aminoácidos polares y cargados y pobre en residuos hidrofóbicos y voluminosos (van der Lee, 2014). No obstante, pueden adquirir una estructura al ser modificadas postraduccionalmente o al unirse a otras moléculas. Un ejemplo es la biomineralización de vertebrados, donde la mayoría de las proteínas asociadas con la formación y crecimiento de la HA son IDP (Boskey, 2016). Se ha demostrado que son partícipes del proceso de biomineralización. Existen subclasificaciones de este grupo de moléculas que se encuentran involucradas en la mineralización de fosfato de calcio (Wojtas, 2012):

- SIBLINGs.
- Proteínas de esmalte.
- Estaterina.

Sólo se hará mención del grupo nombrado SIBLINGs.

3.2.1 SIBLINGs

“*Small integrin-binding ligand, N-linked glycoproteins*” o SIBLINGs son un grupo de NCP consideradas como una familia de proteínas compuesta por cinco glicofosfoproteínas de unión a integrinas (Bellahcène, 2008):

- Osteopontina (OPN).
- BSP.
- Proteína de la matriz dentinaria 1 (DMP1).
- Sialofosfoproteína de dentina (DSSP).
- Fosfoglicoproteína de matriz extracelular (MEPE).

El término “SIBLINGs” hace referencia a la familia de genes cuyos productos comparten características bioquímicas y genéticas, mismas que son (Wojtas, 2012):

- I. Organización de genes y localización cromosómica similar.
- II. Existencia de motivos RGD en su secuencia, que regulan la adhesión y señalización celular.
- III. Modificaciones postraduccionales: fosforilaciones y glicosilaciones.
- IV. Abundancia de residuos de aminoácidos ácidos.
- V. Capacidad de unión a colágeno y iones Ca^{2+} , y presencia en tejidos mineralizados.
- VI. Carácter molecular intrínsecamente desordenado.

De este grupo es de particular interés la BSP.

3.2.2 BSP

La BSP fue reportada por primera vez en 1972, sin embargo, sólo se trataba de un fragmento de la proteína. La BSP completa se purificó a partir de hueso bovino, se describió como una proteína con una masa molecular relativa (M_r) de 59,000, un alto contenido de ácido siálico y rica en oligosacáridos, punto isoeléctrico de 3.9 y alto contenido de glutamato y aspartato (Heinegård, 1989). En general, BSP es una glicoproteína de ~68 kDa (Kim, 1994) codificada por el gen IBSP localizado en el

cromosoma 4 en humanos (Fisher, 1990). Su estructura primaria consta de 317 aminoácidos (Anexo 2), de los cuales los 16 primeros aminoácidos corresponden a la péptido señal. La proteína se divide en tres dominios: dominios N y C terminal, residuos 1-45 y 219-301 respectivamente, y un dominio central, residuos 46-218 (Ganss, 1999). BSP contiene una secuencia de reconocimiento de integrinas, un dominio RGD (arginina-glicina-aspartato), el cual es crítico para la señalización celular y en la diferenciación de los osteoblastos (Baht, 2008). Se ha mostrado que BSP puede unirse a distintos ligandos como colágeno, metaloproteinasas, integrinas e hidroxapatita (Gordon, 2007). Además, mediante experimentos realizados con $^{45}\text{Ca}^{2+}$ se demostró la capacidad de interacción de BSP con iones Ca^{2+} , cuyos resultados indican una unión a 83 iones Ca^{2+} por molécula y una constante de unión de 0.5-0.1mM (Chen, 1992). Mediante una modificación del método de Scatchard y usando BSP, purificada de una línea celular de osteosarcoma de rata, se calculó una $K_d \approx 2.6 \times 10^{-9}\text{M}$ para interacción BSP con HA (Stubbs, 1997). Adicionalmente, se demostró mediante experimentos de cuantificación por absorción atómica la pérdida de Ca^{2+} en el sobrenadante de distintas soluciones con fragmentos de BSP, corroborando la capacidad de BSP de secuestrar iones Ca^{2+} de la solución (Stubbs, 1997). También, se ha observado el potencial de BSP con modificaciones postraduccionales para inhibir la nucleación de HA (Stubbs, 1997). A pesar de estos datos, Hunter (1993) demostró que BSP tiene la facultad de nuclear cristales de HA, utilizando BSP purificada de la calvaria de cerdos adultos. Las IDP de los tejidos mineralizados, pueden actuar de manera distinta en la nucleación de los minerales debido a la posición de las modificaciones postraduccionales, la concentración de la proteína y su origen (Villarreal-Ramirez bone 2017).

BSP en mamíferos tiene una conservación en secuencia del 45% y entre un 10% y 23% de sustituciones conservativas (Ganss, 1999). Las regiones con mayor conservación en la secuencia corresponden a los extremos amino y carboxilo terminal, la secuencia RGD y las regiones ricas en ácido glutámico. Es generalmente aceptado, que una secuencia conservada filogenéticamente suele

implicar su importancia biológica funcional y estructural (Ponting, 2017). Estudios de la estructura de BSP usando NMR unidimensional concluyeron que BSP tiene una estructura flexible sin regiones estructurales significantes (Fisher, 2001). Los resultados de dicroísmo circular (CD) reportados por Wuttke (2013) indican un contenido de 5% hélice α , 32% lámina β , 17% giros β y 46% giros aleatorios. Mientras, los estudios reportados por Tye (2003) mostraron en CD para BSP recombinante un contenido en estructura secundaria de 13.8% de hélice α , 43.5% de lámina β , 10.8% de giros β , 8.1% de hélice poli-pro (II) y 18.9% de estructura desordenada. Además, en estudios de dispersión de rayos X de bajo ángulo (SAXS) se observó una estructura de giros aleatorios (Tye, 2003).

BSP es una proteína con una carga global negativa y con un alto grado de fosforilación. Experimentalmente, se ha demostrado que BSP tiene, de manera natural (*in vivo*), aproximadamente 5.8 mol de serina fosforilada por mol de proteína, modificaciones correspondientes a las enzimas CKII, PKC y cinasa dependiente de cGMP (Salih, 2003). Estudios realizados en BSP bovina muestran que hay 11 sitios potenciales identificados para ser fosforilados, sin embargo, no todos se encuentran fosforilados para una población dada de moléculas aisladas de hueso a una edad específica, mostrando así una heterogeneidad en las modificaciones postraduccionales (George, 2008). Además, se sabe que la mayoría de estas fosforilaciones ocurren en la región cercana al extremo amino terminal. Zaia (2001) reporta las siguientes fosforilaciones encontradas de manera experimental mediante MALDI-TOF:

- Fosforilación en la Ser-15.
- Fosforilaciones heterogéneas entre las posiciones 130-203 con número de entre 0 y 4 posibles fosforilaciones, los residuos más probables para estas modificaciones son, Ser-133, 177 y 189, Thr-167.

Además, sugiere que la presencia de estas fosforilaciones y los niveles de fosfatasa alcalina en la matriz orgánica pueden ser un mecanismo de regulación para la actividad de la BSP. Este clúster de fosforilaciones puede influenciar la capacidad

de unión a HA de dos zonas de repeticiones de ácido glutámico debido a la alta carga negativa (Ganss, 1999).

En esta tesis se busca identificar segmentos de la BSP cuyas propiedades sean de interés para estudiar, mediante simulaciones de DM, las interacciones con iones Ca^{2+} así como el efecto que tiene la presencia de fosforilaciones en estas interacciones y en características estructurales de los distintos péptidos en estudio. Los criterios por considerar en los fragmentos de BSP son los resultados de la predicción de IDRs, de fosforilaciones y estructura secundaria, y conservación de secuencia, presencia de cargas eléctricas, así como la información existente en bibliografía. Se espera que algunos de los péptidos que reúnan las características ya mencionadas muestren una fuerte afinidad hacia el Ca^{2+} y que existan cambios conformacionales tras la interacción entre péptido y iones, siendo así potenciales dominios funcionales de la BSP. De esta manera se busca ampliar el conocimiento existente acerca de BSP y su rol funcional en la biomineralización. Este proyecto pretende, a futuro, realizar estudios de asociación/interacción con otros ligandos fisiológicos de BSP (HA, colágeno), así como la verificación experimental de los resultados obtenidos en este trabajo.

4.0 Metodología y bases teóricas

Para evaluar el papel de las fosforilaciones en BSP y su participación en la unión a los iones Ca^{2+} se utilizaron distintas herramientas bioinformáticas y simulaciones de DM. A continuación, se describen los fundamentos de las herramientas y procedimientos realizados. Las nomenclaturas utilizadas para nombrar a los aminoácidos se encuentran en el Anexo 1.

4.1 Bioinformática

La bioinformática es una ciencia multidisciplinaria que incluye el conocimiento de varias disciplinas como matemáticas, física, medicina y biología, cuyo principal propósito es la captura, almacenamiento, análisis e interpretación de datos para la solución de problemas biológicos (Bayat, 2002). Históricamente se considera a Margaret O. Dayhoff como madre de la bioinformática, debido a su papel en el desarrollo de matrices para el análisis de secuencias de nucleótidos y aminoácidos (Hunt, 1984). La bioinformática ha encontrado soluciones en la predicción de estructuras tridimensionales de macromoléculas biológicas, dilucidación de mecanismos de catálisis enzimática y en el desarrollo de fármacos (Xu, 2015), por mencionar algunas. Se han desarrollado un gran número de algoritmos de análisis, comparación y predicción, como lo son el BLAST y el aprendizaje automatizado o “*machine learning*” (ML), los cuales fueron de utilidad en el desarrollo del presente trabajo.

4.1.1 BLAST

“*Basic Local Alignment Search Tool*” o BLAST, por sus siglas en inglés, es un algoritmo creado en 1990 por un grupo de trabajo en el *National Center for Biotechnology Information* (Altschul, 1990). Se trata de un programa para búsqueda de similitudes de secuencia en bases de datos de proteínas y ácidos nucleicos. Entre sus distintos usos se puede mencionar la búsqueda de filogenias, homologías, identificación de especies, entre otros. BLAST es un heurístico que compara secuencias dadas por el usuario con secuencias en bases de datos y calcula la

significancia estadística de aquellas que coincidan. El algoritmo se puede describir de la siguiente manera (Altschul, 1990,1997; Mount, 2001):

- De manera opcional, la secuencia se puede filtrar para remover regiones de baja complejidad (LCR, por sus siglas en inglés).

Una LCR es aquella zona de una proteína que tiene una baja diversidad en su composición de aminoácidos. Esto puede ser: zonas que contengan muy pocos aminoácidos diferentes o bien zonas que contienen un solo aminoácido. Las posiciones pueden encontrarse como clúster (agrupadas o en cúmulos) periódicamente o de manera irregular, lo que implica un bajo contenido de información útil para análisis computacional (Coletta, 2010).

- Se genera una lista de palabras (ventanas) de longitud w a partir de la secuencia de consulta.

Las palabras están formadas por un cierto número de letras que representan aminoácidos o nucleótidos, según sea el caso. Su longitud dependerá del tipo de molécula en estudio siendo $w=3$ para péptidos y $w=11$ para DNA. Las palabras se forman para cada una de las posiciones desde el inicio de la secuencia hasta llegar a los últimos tres aminoácidos.

- Se agrupan las palabras y se evalúan con la matriz de puntuaciones BLOSUM62 para una coincidencia exacta. También se evalúan las posibles coincidencias con cualquier combinación similar de aminoácidos.

Una matriz de puntuación o “*scoring matrix*” es una herramienta que se utiliza para estimar la similitud de un alineamiento de secuencias. BLOSUM62 o “*Blocks of AminoAcid Substitution Matrix 62*” es una matriz que fue calculada a partir de alineamientos con un máximo de 62% de identidad (Henikoff, 1992). En la Figura 2 se muestran los valores de la matriz para cada par de aminoácidos

$EDM = 5 + 6 + 0 = 11$. Cuando se tienen los valores para las palabras generadas, el valor T ayuda a reducir el número de posibles coincidencias para la generada a partir de la secuencia. Sólo aquellas palabras por encima de T se conservan para la búsqueda. Suponiendo un $T=10$ para el ejemplo anterior, únicamente “EDM” sería considerada para los siguientes pasos.

- El procedimiento anterior se repite con cada palabra generada a partir de la secuencia inicial.
- Se escanean las secuencias presentes en la base de datos en busca de coincidencias exactas con las palabras elegidas para búsqueda.
- El siguiente paso tiene como propósito extender las coincidencias entre las secuencias para encontrar un segmento de secuencia más grande, conocido como “*high-scoring segment pair*” o HSP, que tenga una puntuación mayor a la de la palabra original. En este punto es necesario hacer una aclaración. La versión de BLAST utilizada actualmente, conocida como BLAST2 o *gapped* BLAST, difiere en este paso del método empleado originalmente (Altschul, 1997). Anteriormente, después de encontrar una coincidencia entre las secuencias se extendía el alineamiento en ambas direcciones de ésta hasta que la puntuación del HSP empezara a disminuir, sin embargo, este paso demanda de mucho tiempo de computo. En la versión actual de BLAST se utiliza un valor T más bajo que el del método anterior, generando así una lista de palabras más grande, a continuación, las coincidencias son representadas como se muestra en la figura 3.

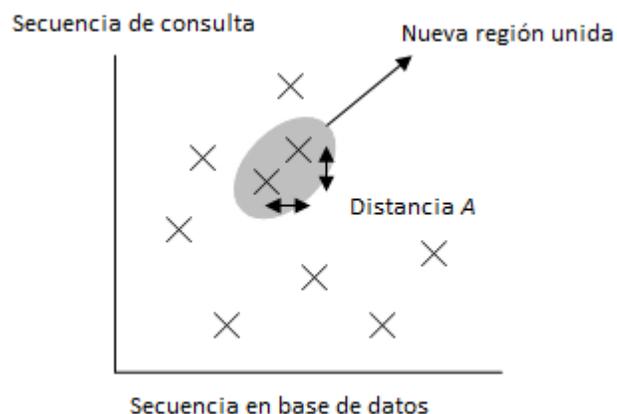


Figura 3. Metodología empleada en BLAST2

(Tomada de la Wikipedia china y modificada para su uso, liberada al dominio público por su autor "DISP").

Las "X" representan las palabras, correspondientes al nuevo valor de T , con coincidencia encontrada. El programa hace uso de aquellas palabras que se encuentren en diagonal y a una distancia A (zona oscura) y procede a unir las para formar un segmento de mayor longitud. Estos nuevos segmentos son utilizados para encontrar los HSP bajo el mismo protocolo que se utilizaba en el BLAST original. El *score* de los HSP encontrados se calcula a partir de la matriz.

- Todos los HSP obtenidos se enlistan y se compara su puntuación con un valor conocido como " S ". Este parámetro es una medida de la similitud que existe entre secuencias comparadas. El valor límite que se utiliza en el algoritmo es determinado de manera empírica examinando una serie de puntuaciones de secuencias aleatorias y seleccionando un valor significativamente mayor. Sólo se conservan aquellos que sean mayores.
- BLAST determina la significancia estadística para el valor de cada HSP.
- Si se encontraron dos o más HSP se puede realizar un alineamiento más largo, dependiendo de sus significancias calculadas previamente.

- En la versión actual de BLAST se producen alineamientos que incluyen *gaps*, sin embargo, estos conllevan penalizaciones para la puntuación final del alineamiento. La puntuación de un alineamiento se calcula tal y como se muestra en la ecuación (2). M es la puntuación de la matriz BLOSUM62 para dos aminoácidos, “ i ” y “ j ”, “ c ” es el número de *gaps* presentes en el alineamiento, “ O ” es la penalización por la existencia de un *gap*, “ d ” es la longitud total de los *gaps* y “ G ” es la penalización por extensión del *gap*. Los valores por defecto son: $O= 11$, $G= 1$ (Kerfeld, 2011).

$$S = \left(\sum M_{ij} \right) - cO - dG \quad (2)$$

- Una vez que se obtiene la puntuación de los alineamientos, estos valores son convertidos a un “valor E ”. Este parámetro describe el número de coincidencias, para un alineamiento, que se pueden esperar por mera probabilidad al buscar en una base de datos. La relación entre la puntuación “ S ” y el valor “ E ” es exponencialmente inversa. Entre más bajo sea su valor mayor significancia tendrán los alineamientos. Este número es comparado con un valor E predeterminado por el algoritmo y aquellos alineamientos que presenten un E mayor se reportan en los resultados.

4.1.2 Aprendizaje automatizado (ML)

Históricamente el ML nace en la década de 1950 con el trabajo de Samuel (1959), donde un programa de computadora para jugar damas tenía la facultad de aprender a la vez que era ejecutado. El aprendizaje automatizado se refiere al proceso donde diversos algoritmos aprenden iterativamente de distintos patrones u observaciones de forma continua para ser capaces de dar predicciones de un grupo de datos (Lai, 2019). En otras palabras, son sistemas capaces de aprender de manera autónoma basándose en observaciones y análisis de patrones de un conjunto de datos.

En el área de la inteligencia artificial hay distintas clasificaciones para la forma en que las máquinas adquieren conocimiento, las más mencionadas en la literatura son las siguientes (Ayodele, 2010):

- ML supervisado.
- ML no supervisado.

El aprendizaje no supervisado es aquel en el que los datos se presentan de manera cruda, es decir, sin ningún tipo de etiqueta o identificador. Su propósito es encontrar patrones en la información y agrupar las muestras basándose en similitudes entre las variables.

Por otro lado, el aprendizaje supervisado es una técnica en la que se entrena al sistema mediante series de datos con etiquetas. Este tipo de ML es punto de interés para los fines de esta tesis. Las etiquetas se presentan en pares a manera de vectores, es decir, especificando un *input* u objeto de entrada y su respectivo *output* o valor de salida deseado. Su uso se puede clasificar de dos maneras dependiendo de la naturaleza del *output*: de clasificación para etiquetas o grupos y de regresión para valores continuos (Ji, 2014). Éste es uno de los sistemas más empleados, por lo tanto, se han desarrollado un gran número de técnicas, algunas de ellas son:

- Redes neurales artificiales (NN).
- Máquinas de vectores de soporte (SVM).
- Clasificación Bayesiana.
- Bosques aleatorios (RF).

Sólo se hablará de NN y SVM, ya que son los algoritmos que conforman los predictores empleados.

4.1.3 Redes neurales artificiales (NN)

Las NN son una técnica de modelado basada en el procesamiento de información que ocurre en el sistema nervioso (Rumelhart, 1986a). Son utilizadas en distintas áreas de la ciencia como herramientas para reconocimiento y clasificación de patrones, identificación y modelado de sistemas, procesamiento de señales e

incluso para análisis de secuencias peptídicas (Wu, 1997). Una de las ventajas de las NN ante métodos estadísticos clásicos como regresión y ANOVA, es que pueden ajustar datos donde la relación entre variables dependientes e independientes no es lineal y donde la forma específica de la relación no lineal es desconocida (Francis, 2001). Las NN se clasifican de acuerdo con su arquitectura, dividiéndose en dos grandes grupos “*Feed Forward*” y “*Feedback*”, la principal diferencia radica en el sentido del flujo de la información: en las redes *feed forward* el flujo es lineal, mientras que las redes de tipo *feedback* hay retroalimentación entre los componentes del sistema (Gardner, 1998). Únicamente el primer tipo es de interés para esta tesis.

Las redes neurales son una técnica que se basa en la idea del perceptrón, también conocido como “*Single Layer Perceptron*” (SLP) (Rosenblatt, 1958), siendo éste el modelo más simple de una red neural. Es de naturaleza binaria y únicamente es capaz de clasificar datos que se puedan separar linealmente. Su arquitectura consiste en dos capas una de entrada y otra de salida, y se puede representar de la siguiente manera:

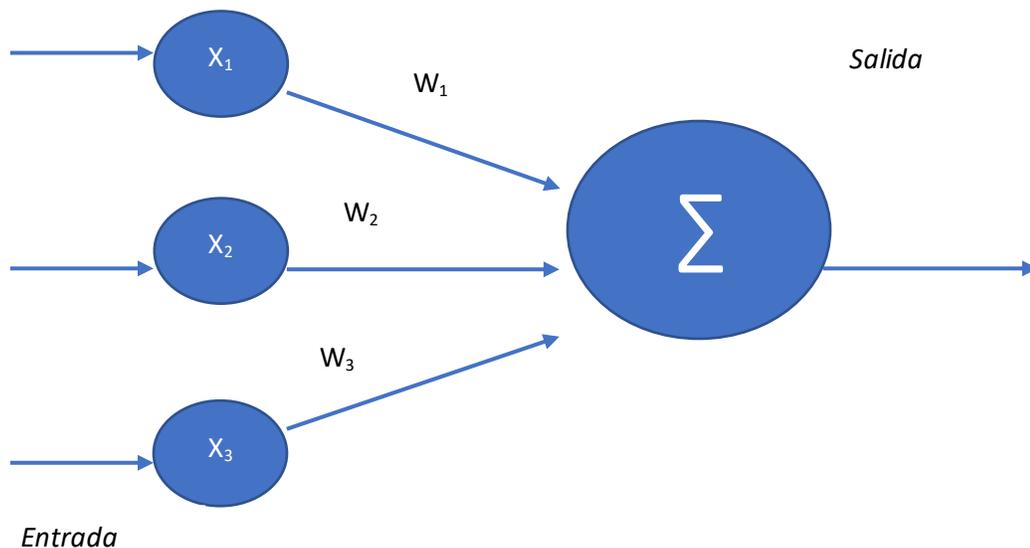


Figura 4. Arquitectura de un perceptrón

El sistema descrito en la figura 4 considera al modelo básico de un perceptrón. Se trata de un sistema que acepta N número de elementos X de entrada, a los que se les asigna un peso w y posteriormente una neurona realiza el computo de una suma ponderada. Basándose en la función escalón de Heaviside (Reber, 2001) los resultados de la suma son comparados con un valor límite θ , y se les asigna un valor binario de acuerdo con la ecuación (3):

$$f(x) = \begin{cases} 1, & \sum_{i=1}^N w_i x_i \geq \theta \\ 0, & \sum_{i=1}^N w_i x_i < \theta \end{cases} \quad (3)$$

Las redes neurales se utilizan en casos donde la relación no es lineal, por lo tanto un modelo SLP no sería suficiente para llevar a cabo estas tareas, es por esto por lo que se utilizan redes neurales de tipo “*Multi Layer Perceptron*” (MLP). Una NN de tipo MLP puede constar de tres o más capas, dependiendo de su complejidad, siendo éstas : una entrada, una salida y el resto se denomina como “*hidden layers*” o capas ocultas.

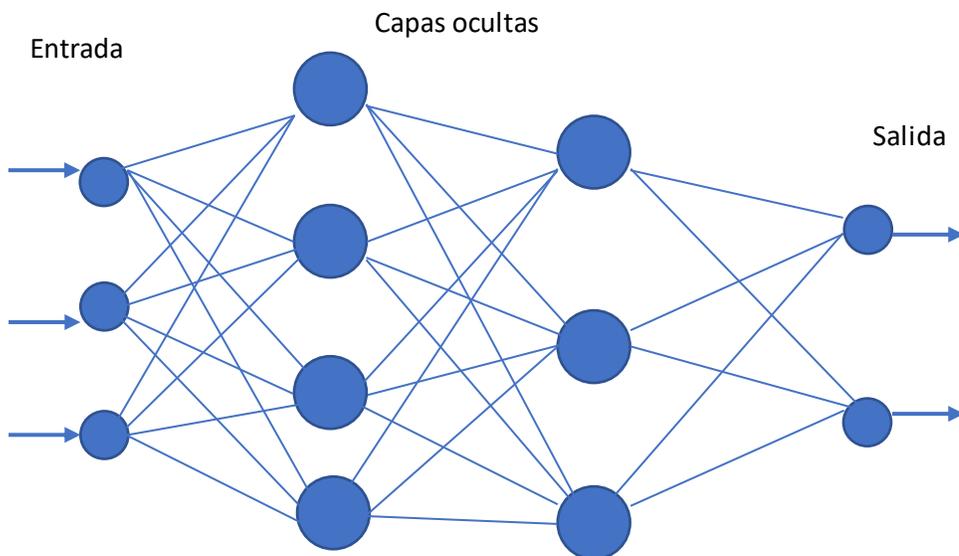


Figura 5. Modelo de un MLP de 4 capas

Como se observa en la Figura 5 una NN de tipo MLP es un sistema en el cual las neuronas se encuentran conectadas entre sí, esto representa la relación no lineal que existe entre el vector de entrada y el de salida. Las señales de salida dadas por cada neurona son la suma de los *inputs* modificada por una función de activación o transferencia (Gardner, 1998). Las NN tienen la habilidad de aprender mediante entrenamiento que consiste, típicamente, en presentarle al MLP una serie de valores de entrada y sus vectores de salida asociados. Éste se realiza mediante un algoritmo conocido como “*backpropagation*” (BP) o retropropagación (Rumelhart, 1986b), que consiste en lo siguiente:

1. Se asignan valores aleatorios a los pesos w de cada conexión.
2. Se presentan los vectores de entrada, obtenidos de los datos de entrenamiento.
3. Propagar el vector de entrada para obtener el valor de salida correspondiente.
4. Se calcula el error de la señal, lo que se realiza comparando el *output* obtenido con el deseado.
5. Propagar el error a través de la red.
6. Se ajustan los pesos w para minimizar el error general.
7. Se repiten los pasos anteriores con los siguientes vectores hasta encontrar un valor de error mínimo satisfactorio.

El principal propósito del algoritmo es el ajuste de los valores w , que puede ser mucho más complejo que los pasos anteriores dependiendo del escenario. Este procedimiento se conoce como descenso de gradiente y busca encontrar el error mínimo absoluto o global, para detalles a mayor profundidad y más contexto matemático consultar (Bishop, 1995).

4.1.4 Máquinas de vectores de soporte (SVM)

Son un tipo de ML (sección 4.1.1) construidas por Cortes y colaboradores, cuyos resultados son de naturaleza binaria (Cortes, 1995) y que surge a manera de alternativa a distintos inconvenientes presentados por las NN. Tal es el caso de

obtener soluciones óptimas a nivel local y no a nivel global, y la excesiva cantidad de datos requeridos para entrenarlas. Hay tres características que las destacan: su habilidad de aprender con un número muy pequeño de muestras, robustez ante los errores del modelo y su eficiencia computacional comparada con métodos como las NN (Gholami, 2017). Es mayormente usado para problemas que involucren clasificación (sección 4.1.1).

En este algoritmo cada dato es representado como un punto en un espacio n -dimensional, donde “ n ” es el número de características. Cada característica está representada por una coordenada. El objetivo es encontrar un hiperplano, también conocido como función de decisión lineal, que permita diferenciar de la mejor manera posible entre las dos clases presentes. Las SVM trabajan seleccionando un hiperplano que equidiste a los datos más cercanos de cada clase, a estos datos se les conoce como “vectores de soporte”. Es decir, se busca que exista un “margen máximo” (Maglogiannis, 2007). Se considera como margen a la distancia que existe entre el hiperplano y los vectores de soporte. A continuación, se explica de manera gráfica, lo expuesto anteriormente.

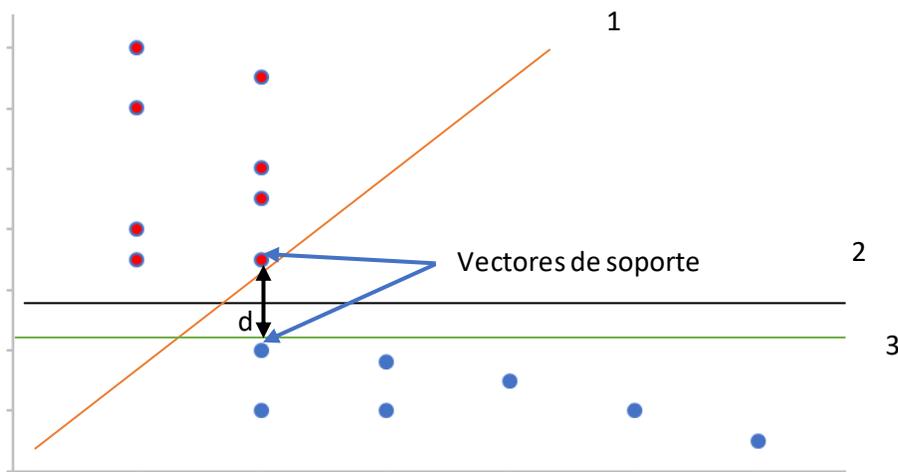


Figura 6. Caso de tres posibles hiperplanos separando dos clases de datos

En la figura 6 se muestran dos clases de datos representadas por colores diferentes, donde se señalan los vectores de soporte para cada grupo y a la vez se aprecian tres hiperplanos distintos (1-3) que pueden ser propuestos como solución. Aunque

cumplen con el propósito de separar las clases, no todos satisfacen el principio de “márgenes máximos” bajo el cual operan las SVM. A primera vista se observa que los hiperplanos 1 y 3 no cumplen con éste, ya que no son equidistantes a los vectores señalados (para fines prácticos la imagen fue construida intencionalmente de esta manera), mientras que el hiperplano 2 mantiene una distancia “ d ” de ambos, siendo así el hiperplano que da solución al problema expuesto.

Sin embargo, lo anterior es la representación de un problema meramente lineal y no es aplicable a los problemas que se observan en la vida real. Para situaciones en las que se presentan clases que no son separables de manera lineal es necesario utilizar espacios de alta dimensionalidad, también conocidos como espacios transformados, y buscar ahí hiperplanos de separación lineal mediante funciones *kernel*. Una función *kernel* es, por definición, una función que asigna a cada par de elementos del espacio de entrada un valor correspondiente al producto escalar de sus vectores en un nuevo espacio. Existen varios tipos como: *kernel* lineal, polinómico, gaussiano y sigmoidal (Suárez, 2016). No se entrará a mayores especificaciones matemáticas (ver sección 4.1.1) y se procede a explicar.

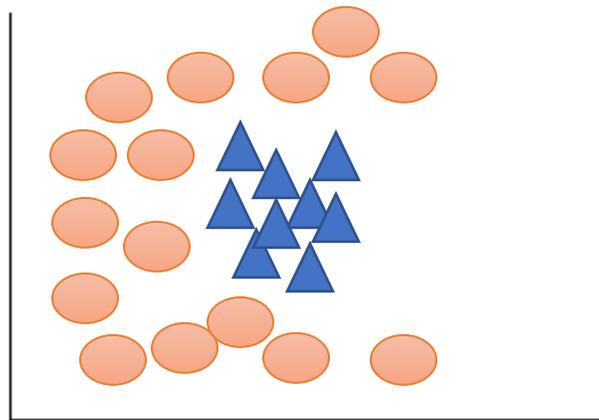


Figura 7. Dos clases de datos que no pueden ser separadas linealmente por un hiperplano

En la figura 7 se muestran puntos que no pueden ser separados de manera adecuada por un hiperplano, por tanto, el algoritmo hace uso de las funciones *kernel*

para producir un espacio transformado en el cual exista una función de decisión lineal, tal y como se muestra a continuación:

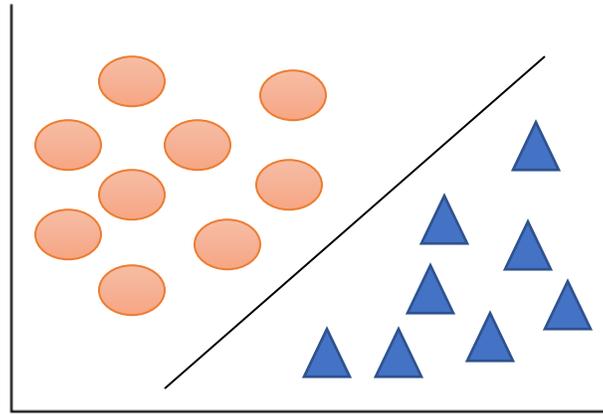


Figura 8. Función de decisión lineal encontrada para un grupo de datos proyectados en un espacio transformado

En la figura 8 se observa el grupo de datos proyectados en un espacio transformado por una función *kernel*. La elección del hiperplano debe cumplir con los criterios explicados al inicio de esta sección. Cuando se encuentra una función que satisfaga los principios ya mencionados, los datos y la función de decisión se regresan al espacio original (Figura 7), obteniendo un plano no lineal que es capaz de realizar la clasificación binaria, tal y como se muestra en la figura 9.

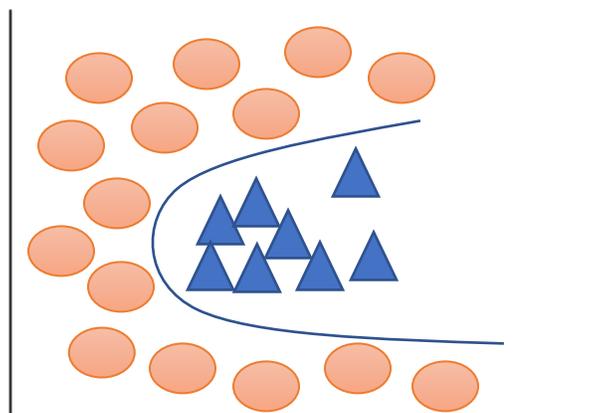


Figura 9. Grupos de datos clasificados en su espacio original

Lo anterior sirve para explicar, de manera gráfica y sencilla, el funcionamiento de una SVM, aunque los problemas reales pueden presentar dimensionalidades más complejas que lo aquí expuesto (Maglogiannis, 2007).

4.1.5 Predicción de fosforilaciones

La adición de grupos fosfato a un péptido es una modificación postraduccional que en mamíferos ocurre, principalmente, en residuos de serina, tirosina y treonina (Stegg, 2003). Se sabe que la fosforilación de una proteína tiene efectos sobre procesos celulares básicos como crecimiento y diferenciación celular, metabolismo, motilidad, transporte membranal, entre otros, y se estima que, aproximadamente, un 30% de las proteínas de una célula se encuentran fosforiladas a la vez (Cieśla, 2011). Tal importancia, aunada a la dificultad experimental que conlleva realizar una identificación de sitios de fosforilación a larga escala estimuló el desarrollo de herramientas computacionales que permitieran predecir la presencia de estos sitios a partir de la secuencia peptídica (Iakoucheva, 2004). Una de las herramientas bioinformáticas de predicción más ampliamente usada, y la elegida para el presente proyecto, es NetPhos 3.1. NetPhos es un servidor web que identifica posibles sitios de fosforilación para proteínas eucarióticas usando ensambles de redes neurales artificiales, combinando dos predictores: Netphos 2.0 que realiza predicciones genéricas (Blom, 1999) y NetphosK 1.0 predice fosforilaciones específicas de 17 cinasas diferentes. (Blom, 2004). Ambos predictores utilizan las características descritas en trabajos anteriores (Blom, 1996). Siendo éstas: una arquitectura del tipo *feedforward* (ver sección 4.1.2) con una capa escondida de neuronas y como función de transferencia la función sigmoide.

El entrenamiento de ambos predictores fue realizado con sitios de fosforilación verificados experimentalmente obtenidos de PhosphoBase, actualmente accesible en <http://phospho.elm.eu.org/>. Para el predictor genérico se utilizaron 584 sitios de serinas, 108 de treoninas y 210 de tirosina. En el caso del predictor cinasa-específico el artículo original sólo hace mención de seis cinasas para serina y treonina (Blom, 2004): PKA, PKC, PKG, cdc2, CK-2 y CaM-II, utilizando para cada

una de ellas entre 22 y 258 sitios diferentes. El único criterio de selección, para ambos predictores, es que las secuencias fueran diferentes en una ventana de 9 aminoácidos a partir del residuo aceptor de grupo fosfato.

El servicio de predicción es accesible mediante el sitio web: <http://www.cbs.dtu.dk/services/NetPhos/>, donde sólo se requiere de enviar la secuencia de interés en formato FASTA y seleccionar si se desea la predicción de un aminoácido en específico o de los tres. Los resultados incluyen la puntuación obtenida (0 – 1), los aminoácidos con una puntuación ≥ 0.5 son reportados como predichos para presentar fosforilación.

Para más detalle acerca del proceso de validación del servidor, sensibilidad y especificidad de éste y de los siguientes métodos de predicción consultar el Anexo 3.

4.1.6 Predicción de IDR

Con el paso de los años el estudio de las IDP se ha vuelto un área que crece rápidamente, esto es debido a estudios que han demostrado su papel en procesos biológicos como control de ciclo celular, regulación, reconocimiento y señalización. Al mismo tiempo se ha detectado una sobreexpresión de IDP en las vías de señalización de procesos patológicos, siendo así un posible blanco para inhibición (Vojisavljevic, 2016). Lo anterior ha dado lugar al desarrollo de herramientas computacionales que permitan la detección de IDR, pues cada vez se les asocian más roles funcionales de importancia (Vullo, 2006).

Para este proyecto se eligieron dos servicios de predicción: PONDR y PrDOS. PONDR (*“Predictors of Natural Disordered Regions”*) integra tres predictores en su algoritmo de trabajo “VL-XT”:

- Predictor VL1 (Romero, 1997): Optimizado para la predicción de zonas largas de desorden o LDR (*“Long Disordered Regions”*), de más de 39 aminoácidos en zonas internas, es decir no realiza predicción sobre las zonas terminales. Basado en redes neurales, su arquitectura consta de una

capa *input*, una capa de *output* y una *hidden layer*, específicamente: 10 unidades de entrada, 10 de procesamiento y una de salida. Entrenado mediante el algoritmo de retro-propagación (ver sección 4.1.2). Originalmente, se desarrolló con proteínas de la familia de la calcineurina.

- Predictores XC y XN, en conjunto llamados XT (Li, 1999). Ambos basados en redes neurales. Entrenados con datos de cristalografía de rayos X, donde las regiones desordenadas constan de 5 o más aminoácidos. Únicamente realizan predicción sobre zonas terminales. La arquitectura de consta de 8 unidades de entrada, 8 unidades de procedimiento y una de salida.

La integración de los predictores se explica brevemente a continuación:

1. Se realizan predicciones en las áreas correspondientes: once aminoácidos antes y después de los extremos carboxilo y amino, respectivamente, para VL1. Para XT las predicciones constan de catorce aminoácidos desde cada término. Esto ocasiona un traslape de predicciones para los aminoácidos 11-14 para el extremo amino terminal y las posiciones X-11 a X-14, considerando a X como la última posición del péptido.
2. Se obtiene la media de los valores traslapados.
3. Se realiza un suavizado de los valores mediante la obtención de medias para grupos de 9 aminoácidos, se excluyen las primeras y últimas 4 posiciones.

El servicio es accesible mediante un servidor web <http://www.pondr.com/> donde sólo se requiere de la secuencia en formato FASTA. El resultado final consta de un número entre 0 y 1, mientras más cercana sea la predicción sea a 1 indica una predicción ideal de desorden, así como de la clasificación, ordenado o no, asignada a cada aminoácido.

PrDOS se divide en dos partes, una basada en el algoritmo SVM y la otra consiste en una búsqueda PSI-BLAST (Ishida, 2007). Anteriormente se ha explicado el funcionamiento de las SVM (ver sección 4.1.3), por lo que a continuación únicamente se explica el funcionamiento de PSI-BLAST. PSI-BLAST o "*Position-Specific Iterative BLAST*" es un algoritmo basado en BLAST2 (ver sección 4.1.4

BLAST), el cual incluye pequeñas modificaciones que le permiten tener una mayor sensibilidad a pequeñas similitudes de secuencia que pueden ser biológicamente relevantes (Schaffer, 2001). A continuación, se explica su flujo de trabajo (Bhagwat, 2007):

- La primera iteración de un PSI-BLAST es idéntica a la realizada por un *gapped* BLAST.
- Genera un alineamiento múltiple de los HSP obtenidos que se encuentren por encima de un valor predeterminado y se calcula un perfil o matriz de puntuaciones de posiciones específicas (PSSM).

Una PSSM es un tipo de matriz de puntuaciones en la que la sustitución de un aminoácido tendrá un valor dependiendo de la posición en la que ocurra, es decir, no será asignado para una sustitución de aminoácidos en la posición X que si la misma sustitución ocurriera en una posición Y, también la frecuencia observada del aminoácido influye en la puntuación (Koua., 2013), a diferencia de una matriz BLOSUM que es independiente de la posición en la que ocurra la sustitución y siempre tendrá la misma puntuación.

- La PSSM generada contiene el patrón de conservación en el alineamiento. Posiciones altamente conservadas reciben un puntaje alto, mientras que posiciones débilmente conservadas reciben un puntaje cercano a cero.
- El perfil generado (la matriz) se utiliza para realizar búsquedas en la base de datos para detectar secuencias que coincidan con los patrones especificados en la PSSM.
- Las nuevas secuencias detectadas con un *score* por encima del especificado sirven para refinar el perfil generado en pasos anteriores. El proceso se repite de manera iterativa hasta cumplir con especificaciones dadas para el número de rondas o hasta llegar a convergencia, por ejemplo, que no haya nuevas secuencias detectadas.

Una vez que se conocen los algoritmos que integran el predictor es factible explicar el flujo general de trabajo, que se divide en cuatro pasos (Ishida, 2007):

- Generar la PSSM mediante iteraciones de PSI-BLAST en las bases de datos en NCBI.
- Predicción por parte del algoritmo SVM usando las PSSM generadas previamente.
- Predicción usando alineamientos con homólogos, cuya estructura ya haya sido determinada, mediante PSI-BLAST. Si no hay resultados significativos, se omite este paso.
- Combinación de resultados de las predicciones que se realiza mediante el cálculo de la media ponderada.

El servicio es accesible mediante su servidor web <http://prdos.hgc.jp>, se ingresa la secuencia en formato FASTA, se asigna un título de proyecto y los resultados se pueden visualizar en interfaz web o por correo electrónico.

4.1.7 Predicción de estructura secundaria

Se hace uso de tres predictores de estructura secundaria: JPred, PSIPRED y CFSSP.

El servidor Jpred actualmente en su versión “Jpred4” (Drozdetskiy, 2015) basa su funcionamiento en el algoritmo Jnet que consiste en redes neurales (Cuff, 2000) y en PSI-BLAST (ver sección 4.1.2 y 4.1.6). El predictor cuenta con dos modos de trabajo, para secuencias individuales y secuencias múltiples, sólo se hablará del primero.

- Se genera un alineamiento múltiple usando PSI-BLAST, se realizan 3 iteraciones. Se remueven aquellas secuencias con una identidad >75% para evitar redundancias, se remueven las regiones que presentan gaps y se genera un perfil el cual sirve de input para las redes neurales.
- Jnet consta de dos redes neurales conectadas, ambas presentan tres capas (*input*, *output* y *hidden layer*). La primera red tiene un *input* de 17 residuos, 9

unidades de procesamiento que forman la *hidden layer* y 3 nodos de salida que dan resultados en grupos de 19 aminoácidos. Esta red predice la propensión de la proteína para formar hélices, láminas y giros en cada posición de la secuencia. La segunda red toma como *input* los grupos generados, cuenta con 9 unidades de procesamiento en la capa oculta y un *output* de tres unidades. Esta última red se encarga de refinar las predicciones hechas para cada posición.

La estructura de redes neurales fue entrenada mediante 480 alineamientos múltiples realizados con PSI-BLAST, cada uno de ellos para una proteína diferente, obtenidos de la versión de 1996 de la base de datos PDB.

El servicio es accesible mediante su servidor web <http://www.compbio.dundee.ac.uk/jpred/>, donde el usuario requiere de ingresar su secuencia, o secuencias múltiples, en formato FASTA. Los resultados son visibles en la interfaz web o vía correo electrónico.

PSIPRED proporciona diversos servicios al usuario, como predicción de función, modelado de dominios y predicción de estructura secundaria (Buchan, 2013). El método de predicción llamado funciona bajo un protocolo que utiliza PSI-BLAST y redes neurales, haciendo uso de tres fases (Jones, 1999), similar al presentado en el servidor Jpred.

- Se obtiene una PSSM a partir de PSI-BLAST, tres iteraciones, para usar como *input* de las redes neurales
- La arquitectura consta de dos redes neurales, donde se realiza una primera predicción y posteriormente una refinación de ésta. La primera red neural consta de quince grupos con veintidós unidades, cada una, como entrada, una *hidden layer* de setenta y cinco unidades y tres unidades de salida. La segunda red neural presenta quince grupos de cuatro unidades, cada una, como *input*, una capa oculta de sesenta unidades y tres unidades de *output* para cada una de las posibles estructuras (hélices, láminas y giros).

El entrenamiento de las NN se llevó a cabo mediante *back propagation* (ver sección 4.12 y 4.1.5). El servicio es accesible mediante su servidor web <http://bioinf.cs.ucl.ac.uk/psipred/>, el usuario necesita la secuencia en formato FASTA y los resultados son mostrados mediante la interfaz web o vía correo electrónico.

Por último, CFSSP es un servicio web de predicción basado en el algoritmo de Chou y Fasman, se encuentra entre los primeros métodos desarrollados para la predicción de estructura secundaria, se basa principalmente en parámetros de probabilidad, de cada aminoácido, para encontrarse en una estructura de tipo secundaria (Kumar, 2013). El método tiene como principio una serie de parámetros conformacionales y frecuencias posicionales determinadas a partir de estructuras peptídicas conocidas (Chou, 1974a,b), que se muestran en la tabla 1.

El algoritmo de trabajo se divide en tres partes esenciales (Kumar, 2013; Prevelige, 1989):

- 1) **Búsqueda de regiones helicoidales:** Cualquier segmento de seis residuos o más con $\langle P_\alpha \rangle \geq 1.03$ así como $\langle P_\alpha \rangle > \langle P_\beta \rangle$ y que cumpla con las siguientes condiciones se predice como helicoidal.
 - Nucleación de la hélice: Se identifican regiones donde cuatro de seis residuos sean formadores (h_α o H_α). Residuos débiles cuentan como 0.5 h_α . La formación de la hélice no es favorable si el segmento contiene 1/3 de aminoácidos interruptores o menos de 1/2 de formadores.
 - Terminación de la hélice: El segmento helicoidal se extiende en ambas direcciones hasta encontrar un tetrapéptido con $\langle P_\alpha \rangle < 1.00$. Hay combinaciones de aminoácidos que pueden detener la propagación de la hélice, como b_4 , b_{3i} , b_{3h} , b_{2i2} , b_{2ih} , b_{2h2} , b_{ih2} e i_4 , donde b, h e i obedecen a la nomenclatura presentada en la tabla X. y también incluyen a B, H e I, el subíndice representa el número de aminoácidos de esa clasificación. Regiones beta adyacentes pueden terminar la hélice.
 - No puede haber prolina en el interior de la hélice o en el extremo C-terminal.

- Fronteras: Pro, Asp y Glu, se encuentran preferentemente en el extremo N-terminal de la hélice. His, Lys y Arg en el extremo C-terminal. Pro, Asp y Arg pueden ser asignados como l_{α} para satisfacer la condición del primer punto.

Tabla 1. Parámetros conformacionales de Chou y Fasman para hélice- α , láminas β y giros β (1974). Extraída y traducida de Kumar (2013)

Residuo	P_{α}	Tipo de α	Residuo	P_{β}	Tipo de β	Residuo	P_t
Glu	1.51	H_{α}	Val	1.70	H_{β}	Asn	1.56
Met	1.45		Ile	1.60		Gy	1.56
Ala	1.42		Tyr	1.47		Pro	1.52
Leu	1.21		Phe	1.38		Asp	1.46
Lys	1.16	h_{α}	Trp	1.37	h_{β}	Ser	1.43
Phe	1.13		Leu	1.30		Cys	1.19
Gln	1.11		Cys	1.19		Tyr	1.14
Trp	1.08		Thr	1.19		Lys	1.01
Ile	1.08		Gln	1.10		Gln	0.98
Val	1.06		Met	1.05		Thr	0.96
Asp	1.01	l_{α}	Arg	0.93	i_{β}	Trp	0.96
His	1.00		Asn	0.89		Arg	0.95
Arg	0.98	i_{α}	His	0.87		His	0.95
Thr	0.83		Ala	0.83		Glu	0.74
Ser	0.77		Ser	0.75	Ala	0.66	
Cys	0.70		Gly	0.75	Met	0.60	
Tyr	0.69	b_{α}	Lys	0.74	b_{β}	Phe	0.60
Asn	0.67		Pro	0.55		Leu	0.59
Pro	0.57	B_{α}	Asp	0.54	B_{β}	Val	0.50
Gly	0.57		Glu	0.37		Ile	0.47

Notación: H_{α} (formador α fuerte), h_{α} (formador α), l_{α} (formador α débil), i_{α} (α indiferente), b_{α} (interruptor α), B_{α} (interruptor fuerte α), H_{β} (formador β fuerte), h_{β}

(formador β), i_β (β indiferente), b_β (interruptor β), B_β (interruptor fuerte β). Información tomada de (Kumar, 2013).

2) Búsqueda de regiones lámina β : Cualquier segmento de cinco residuos o más con $\langle P_\beta \rangle \geq 1.05$ así como $\langle P_\beta \rangle > \langle P_\alpha \rangle$ y que cumplan con las siguientes condiciones son predichas como lámina β .

- Nucleación de lámina β : Se escanea para identificar zonas donde tres de cinco residuos sean formadores (H_β o h_β). La formación no se ve favorecida si el segmento contiene 1/3 de aminoácidos interruptores o menos de 1/2 de formadores.
- Terminación de lámina β : Se extiende la lámina en ambas direcciones, hasta ser terminada por tetrapéptidos con un $\langle P_\beta \rangle < 1.00$. Regiones alfa adyacentes pueden terminar las regiones beta.
- La presencia de Glu en una zona beta es muy rara, así como la de Pro en el interior de una zona beta.
- Fronteras: Residuos cargados no son frecuentes en las regiones N-terminales, y muy poco frecuente en el interior y en el extremo C-terminal. La presencia de Trp se da la mayoría de las veces en el extremo N-terminal y rara vez en el C-terminal.

3) Predicción de giros: Los giros son predichos únicamente bajo el siguiente concepto $\langle P_\beta < P_t > P_\alpha \rangle$.

Este método es accesible mediante un servidor web <http://www.biogem.org/tool/chou-fasman/>, donde el usuario debe dar la secuencia en formato FASTA, para que se realice la predicción, los resultados son presentados en la interfaz web.

4.1.8 Alineamiento múltiple de secuencias (MSA)

Un alineamiento múltiple o MSA (por sus siglas en inglés) es una herramienta básica de la bioinformática que tiene presencia en varios aspectos de los análisis biológicos moleculares, bien puede ser detectar residuos clave en una proteína, predecir

estructuras secundarias de péptidos o inferir la historia evolutiva de una familia proteica (Edgar, 2004).

Actualmente existe un diverso número de paqueterías que realizan MSA, entre ellas MAFFT (“*Multiple Alignment using Fast Fourier Transform*”). Es un programa que integra el uso de la transformada rápida de Fourier (FFT) (Kato, 2002), el cual ha probado ser uno de los métodos más precisos entre programas como Clustal, Muscle, T-COFFE, entre otros (Niu, 2006). Por lo tanto, es el software de elección para este trabajo. MAFFT consta de 3 metodologías para realizar un MSA, donde cada una es una mejora de la anterior y el uso de una u otra dependerá del análisis que se lleve a cabo (Kato, 2008)

- Métodos progresivos usando el algoritmo “*PartTree*”: Utilizados para alineamientos a larga escala; $N > 10000$, donde “N” es el número de secuencias.
- Métodos progresivos de alineamiento: Utilizados para alineamientos a mediana escala; $10000 > N > 200$.
- Métodos de refinamiento iterativo: Utilizado para alineamientos a pequeña escala; $N < 200, L < 10000$, donde “L” es el número de aminoácidos en la secuencia.

Para la finalidad de este trabajo, el último método es el más adecuado, ya que el MSA buscado consta de 26 secuencias de aproximadamente 300 residuos, las cuales se pueden consultar en el Anexo 2: “Secuencias de BSP”. Sin embargo, el método de refinamiento iterativo necesita de un *input* generado por el método progresivo, por lo que a continuación se explican (Kato, 2002, 2005, 2008).

El método progresivo se divide en dos partes conocidas como FFT-NS-1 y FFT-NS-2.

1) FFT-NS-1:

- En primer lugar, se construye, de manera rápida, una matriz de distancias basándose en un algoritmo que identifica el número de

hexámeros idénticos entre dos secuencias, esto se realiza para todas las secuencias problema. En la interfaz web el usuario puede denominar una secuencia para que todo el análisis se base en ésta, para nosotros la secuencia humana.

- Se construye un árbol guía basado en la matriz de distancias. Se emplea el método UPGMA, cuya función es realizar un agrupamiento jerárquico con base en los valores de la matriz previamente construida, para obtener las diferentes ramas del árbol (Hale, 1988).
- Posteriormente, el algoritmo FFT y una matriz de similitud (los cuales no se describen aquí debido a su extensión, para un entendimiento más detallado consultar (Kato,2002)) son empleados para alinear progresivamente las secuencias de acuerdo con las ramificaciones presentadas en el árbol guía. Esto constituye un primer ciclo de alineamiento.

2) FFT-NS-2:

- Corresponde al segundo ciclo del alineamiento progresivo. Aquí se “refina” y calcula una nueva matriz de distancias basadas en el alineamiento obtenido del primer ciclo.
- Se construye un nuevo árbol guía a partir de la matriz calculada previamente.
- Se repite el alineamiento progresivo tal y como se realiza en FFT-NS-1.

A continuación, el programa puede realizar el refinamiento iterativo del MSA obtenido en el punto anterior:

- El MSA obtenido es dividido en dos grupos, de acuerdo con el árbol guía.
- Estos grupos son realineados usando el mismo algoritmo del método progresivo.
- El proceso es repetido hasta que no haya mejoras en la puntuación del MSA.

MAFFT es accesible mediante su servidor web <https://mafft.cbrc.jp/alignment/server/>, donde un archivo en formato FASTA con las secuencias para alinear es requerido, el usuario puede modificar el método de alineamiento o dejar que el servidor decida de manera automática. Es posible indicar al programa si se desea realizar el MSA usando alguna secuencia como referencia. Como se mencionó anteriormente, se modificó esta opción para que todo se realizara en torno a la secuencia de BSP humana. Los resultados son mostrados en la interfaz web.

4.2 Simulaciones de DM

La dinámica molecular es una de las dos principales familias de técnicas empleadas para realizar simulaciones computacionales. Su origen se atribuye los experimentos realizados en el *Lawrence Livermore National Laboratory* en la década de los 50's (Alder,1957), donde moléculas en una fase líquida se representaron como “esferas duras” para estudiar su arreglo. En la actualidad, las simulaciones de DM son utilizadas para describir, estudiar y predecir patrones y propiedades del comportamiento de proteínas, interacciones entre una molécula y un receptor para el diseño de fármacos, los cambios conformacionales que puede sufrir una molécula bajo ciertas condiciones ya sea en bioquímica o en ciencia de materiales, así como cualquier tipo de evento que requiera de la evaluación de propiedades moleculares de un sistema (Vlachakis, 2014).

El principio básico detrás de las simulaciones de DM es la resolución numérica de las ecuaciones de movimiento de Newton para un sistema de átomos que interactúan. Las cuales se representan por la ecuación (4) o por el sistema de ecuaciones (5).

$$\frac{d^2r_i}{dt^2} = \frac{F_i}{m_i} \quad (4)$$

$$\begin{aligned}\frac{dr_i}{dt} &= v_i \\ \frac{dv_i}{dt} &= \frac{F_i}{m_i}\end{aligned}\tag{5}$$

Recordando que la fuerza puede ser expresada en función de la energía potencial como:

$$F_i = -\nabla U_i\tag{6}$$

La resolución del sistema de ecuaciones implica el uso de métodos de integración numérica. La ecuación (6) será retomada en otra sección. El software de elección para realizar las simulaciones de DM es GROMACS (*Groeningen Machine for Chemical Simulations*) (Berendsen, 1995).

4.2.1 Integrador *Leap-frog*

Existen diversos métodos de integración para la resolución de las ecuaciones (4) y (5). El algoritmo por defecto de la paquetería GROMACS es el integrador “*Leap-frog*” o “salto de rana” (Berendsen, 1995). Se trata de una variante del algoritmo de Verlet (Verlet, 1967).

El algoritmo utiliza las siguientes relaciones:

$$v(t + \frac{1}{2}\Delta t) = v(t - \frac{1}{2}\Delta t) + \frac{\Delta t}{m}F(t)\tag{7}$$

$$r(t + \Delta t) = r(t) + \Delta t v(t + \frac{1}{2}\Delta t)\tag{8}$$

Donde r son las posiciones, t es el tiempo, v las velocidades y $F(t)$ las fuerzas. Su funcionamiento se describe de la siguiente manera: en primer lugar las velocidades son calculadas al tiempo $t + \frac{1}{2}\Delta t$, posteriormente se usan para calcular las posiciones al tiempo $t + \Delta t$. Es de aquí donde surge el nombre del algoritmo, pues

se dice que el cálculo de las posiciones “salta” a las velocidades y después el cálculo de las velocidades “saltan a las posiciones”. Esquemáticamente se representa de la siguiente manera:

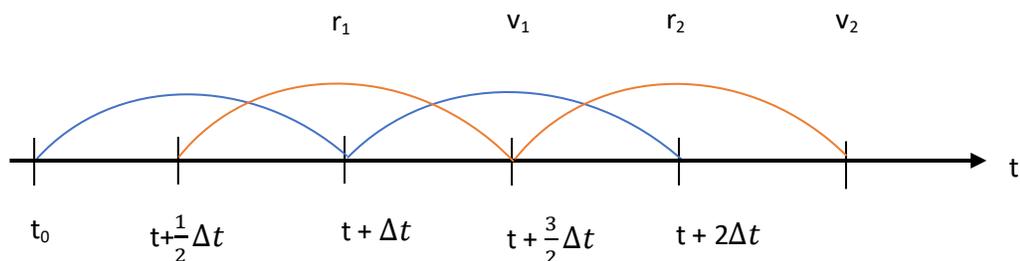


Figura 10. Ilustración del algoritmo *leap-frog* o salto de rana

4.2.2 Campo de fuerza

En la sección 4.2 se habló del principio que rige a las simulaciones de DM, donde se involucra la resolución de las ecuaciones de movimiento de Newton. En la ecuación (6) se define a la fuerza como el gradiente negativo de la función de energía potencial. Esta función está definida por un grupo de funciones de interacción y parámetros conocidos como “campo de fuerza” (Wassenaar, 2006). Se puede definir un campo de fuerza como una expresión matemática que describe la dependencia de energía de un sistema en las coordenadas de sus partículas (González, 2011). El núcleo de cualquier campo de fuerza es la función de energía potencial utilizada para describir la relación de la estructura y la energía del sistema. Una expresión (conjunto de ecuaciones) típica de un campo de fuerza biomolecular se muestra continuación (Mackerell, 2004):

$$U(r) = U_{enlaces} + U_{torsión} + U_{diedro} + U_{Lennard-Jones} + U_{Coulomb} \quad (9)$$

$$U(r) = U_{enlace} + U_{interacciones\ no\ enlazantes} \quad (10)$$

El campo de fuerza de elección es GROMOS96 43a1 con extensión para residuos fosforilados y parámetros HA de Hauptman (Hauptmann, 2003) . La función de energía potencial que utiliza el campo de fuerza se encuentra dada por la siguiente expresión:

Desarrollando la ecuación (10) se obtiene lo siguiente, donde los términos (11)-(14) corresponden a las energías de enlace y (15)-(17) a las interacciones no enlazantes (Daura, 1998, Scott, 1999):

$$U(r) = \frac{1}{4} \sum_{b=1}^{N_b} K_{b_n} [b_n^2(t) - b_{0_n}^2]^2 \quad (11)$$

$$+ \frac{1}{2} \sum_{n=1}^{N_\theta} K_{\theta_n} [\cos(\theta_n(t)) - \cos(\theta_{0_n})]^2 \quad (12)$$

$$+ \frac{1}{2} \sum_{n=1}^{N_\xi} K_{\xi_n} [\xi_n(t) - \xi_{0_n}]^2 \quad (13)$$

$$+ \sum_{n=1}^{N_\phi} K_{\phi_n} [1 + \cos(\delta_n) \cos(m_n \phi_n(t))] \quad (14)$$

$$+ \sum_{\substack{\text{pares no} \\ \text{enlazantes } (ij)}} \left[\frac{C_{12}(ij)}{r_{ij}(t)^6} - C_6(j) \right] \frac{1}{r_{ij}(t)^6} \quad (15)$$

$$+ \sum_{\substack{\text{pares no} \\ \text{enlazantes } (ij)}} \frac{q_i q_j}{4\pi\epsilon_0\epsilon_1} \frac{1}{r_{ij}(t)} \quad (16)$$

$$+ \sum_{\substack{\text{pares no} \\ \text{enlazantes (ij)}}} \frac{q_i q_j}{4\pi\epsilon_0\epsilon_1} \left[-\frac{C_{rf} r_{ij}(t)^2}{2R_{rf}^3} - \frac{1 - 1/2 C_{rf}}{R_{rf}} \right] \quad (17)$$

- La ecuación (11) representa la energía de tensión de los enlaces covalentes del sistema (N_b), donde k_b es la constante de fuerza, b_n y b_{0_n} son las longitudes de enlace instantánea e ideal, respectivamente.
- La ecuación (12), representa la energía de flexión de los ángulos de enlace (N_θ). La constante de fuerza se representa por K_{θ_n} , mientras que θ_n y θ_{0_n} son los valores instantáneos e ideales, respectivamente, para el ángulo de enlace n .
- La ecuación (13) representa la energía potencial asociada a los ángulos diedros impropios (N_ξ). Donde K_{ξ_n} es la constante de fuerza, ξ_n y ξ_{0_n} son el valor instantáneo e ideal del diedro, respectivamente.
- La ecuación (14) representa la energía de torsión de un grupo de ángulos diedros (N_ϕ). La constante de fuerza está representada por K_{ϕ_n} , el cambio de fase está dado δ_n , mientras que m_n es la multiplicidad del sistema.
- La ecuación (15) representa las interacciones de van der Waals entre dos átomos (i y j).
- La ecuación (16) representa las interacciones electrostáticas.
- La ecuación (17) corresponde a la permitividad relativa de una esfera y está dada por el campo de reacción de Poisson-Boltzman.

4.2.3 Minimización de energía

Al construir una molécula con algún software es muy posible que la estructura inicial se encuentre muy distante de su estado de equilibrio. En estas situaciones el computo de fuerzas puede ser excesivo y, por lo tanto, provocaría un fallo en la simulación de DM. En estos casos es necesario realizar un proceso conocido como "Minimización de energía". Por ejemplo, en el arreglo estructural de una proteína puede haber átomos de residuos vecinos que estén demasiado cerca el uno del

otro, lo cual provocaría un aumento en la energía global del sistema, la minimización de energía busca corregir estos problemas mediante optimizaciones donde los átomos son movidos buscando reducir estas interacciones. Se trata de un proceso numérico que busca puntos locales donde la energía es mínima, en comparación con la estructura inicial (Mackay, 1989). Existen distintos algoritmos para realizar esta tarea en GROMACS, sólo se describe el “*steepest descent*” o “descenso más pronunciado”. La siguiente explicación fue extraída de (Abraham, 2018):

En primer lugar, se debe definir un vector como el vector de las coordenadas tridimensionales. Se debe asignar un desplazamiento máximo h_0 , en este caso de 0.01nm. Después las fuerzas y la energía potencial son calculadas. Se calculan nuevas posiciones:

$$r_{n+1} = r_n + \frac{F_n}{\max(|F_n|)} h_n \quad (18)$$

h_n es el desplazamiento máximo y F_n es la fuerza, en este punto es importante recordar la relación entre la fuerza y la energía potencial (ecuación (6)). El denominador $\max|F_n|$ hace referencia a la fuerza escalar más grande en algún átomo. Las fuerzas y las energías se vuelven a calcular para las nuevas posiciones.

- Si ($U_{n+1} < U_n$) las nuevas posiciones son aceptadas y $h_{n+1} = 1.2h_n$.
- Si ($U_{n+1} \geq U_n$) las nuevas posiciones se rechazan y $h_n = 0.2h_n$

El algoritmo se detiene bajo dos escenarios:

1. Se alcanza el número de iteraciones especificadas por el usuario. Para este trabajo: 50000.
2. El máximo de los valores absolutos de los componentes de fuerza es menor a un valor especificado. Para este trabajo: 1000.0 kJ/mol/nm.

Una vez terminado el proceso de minimización de energía se tiene una estructura relajada y energéticamente estable. A continuación, se realiza un proceso de equilibrio entre la molécula en estudio (ya minimizada) y el resto del sistema: agua,

iones, etc. Hasta hace algunas décadas las simulaciones de MD se limitaban a considerar el ensamble micro canónico (NVE) para un sistema, lo que implica un número de partículas (N), volumen (V) y energía total (E) constantes (Heyes, 1983). Sin embargo, además de que puede ser complicado saber el valor energético del sistema, experimentalmente es una condición muy difícil de alcanzar ya que implica la existencia de un sistema perfectamente aislado. Debido a esto es preferible utilizar otros ensambles donde, en principio, cada uno representa una simulación de dinámica molecular, con todo lo que ello implica (ver sección 4.2).

4.2.4 Ensamble NVT

Cuando un sistema simulado se acopla a un baño de calor se forma un ensamble canónico (también conocido como NVT). Este ensamble presenta como características mantener un número de partículas (N), volumen (V) y temperatura (T) constantes. Cuando se tiene interés en llevar el sistema a una temperatura específica es necesario emplear una simulación NVT con un termostato.

Para este trabajo se utiliza el termostato de Berendsen. Se trata de un algoritmo que tiene por principio el “escalado de velocidades” para alterar la temperatura del sistema. Este método se basa en el cálculo del cambio de temperatura ΔT al multiplicar las velocidades en cada paso por un factor λ . Recordando que el ensamble NVT es una simulación de dinámica molecular y las consideraciones descritas en la sección 4.2.1 la ecuación empleada es (Abraham, 2018):

$$\lambda^2 = 1 + \frac{\Delta t}{\tau} \left(\frac{T_0}{T(t - \frac{\Delta t}{2})} - 1 \right) \quad (19)$$

Donde la constante τ representa el acoplamiento entre el sistema y el baño de calor. Una vez que el sistema ha alcanzado la temperatura deseada, se procede con el siguiente ensamble.

4.2.5 Ensamble NPT

También conocido como ensamble isotérmico-isobárico, su principal función será el estabilizar la presión y densidad del sistema, de manera análoga a lo que realiza el ensamble NVT con la temperatura. Las constantes en este proceso de son el número de partículas (N), la presión (P) y la temperatura (T). Es una simulación que parte de un sistema cuya temperatura ha sido estabilizada previamente.

El proceso consta de un termostato (el mismo empleado en el NVT) y un baróstato. La modificación de la presión del sistema puede traer como consecuencia un cambio en las dimensiones de la celda de simulación.

En su expresión matemática, los vectores de la celda se representan por una matriz \mathbf{b}

$$\frac{d\mathbf{b}^2}{dt^2} = V\mathbf{W}^{-1}\mathbf{b}^{l-1}(\mathbf{P} - \mathbf{P}_{ref}) \quad (20)$$

Donde V es el volumen de la caja, \mathbf{W} es una matriz de parámetros que determina la fuerza del acoplamiento. Las matrices \mathbf{P} y \mathbf{P}_{ref} son la presión actual y de referencia, respectivamente.

$$(\mathbf{W}^{-1})_{ij} = \frac{4\pi^2\beta_{ij}}{3\tau_p^2L} \quad (21)$$

El inverso de la matriz de parámetros de masa determina como puede ser deformada la caja, esto es calculado automáticamente por GROMACS, el usuario sólo debe proporcionar las compresibilidades isotérmicas β y la constante de tiempo-presión τ (Abraham, 2018).

Para el análisis de las simulaciones de DM se emplean distintas herramientas de la paquetería GROMACS, las cuales se describen brevemente a continuación.

4.2.6 GROMACS

El análisis de las simulaciones de DM se realiza empleando distintas herramientas integradas en el software. Se obtiene el archivo de trayectoria de cada sistema y se visualizan en PyMOL. Se obtuvo el radio de giro (Rg) para cada péptido. El radio de giro es definido como promedio de la raíz media cuadrática (RMS) de la distancia de todos los elementos que se dispersan del centro de masa de la molécula. El valor del Rg da una idea del empaquetamiento que tiene la molécula. Por ejemplo: comparando dos valores, el valor más bajo indica un empaquetamiento más compacto (Lobanov, 2008). La compactación de una proteína tiene una relación directa con la tasa de plegamiento. (Sneha, 2016). Así como el RMSF y el Factor B, donde el RMSF es la fluctuación que se observa entre residuos o átomos presentes en una macromolécula. La fluctuación atómica puede explicar el nivel de flexibilidad de una proteína durante una simulación. (Sneha, 2016). Mientras que el B-factor, factor de Debye-Waller o factor de temperatura, refleja las fluctuaciones de los átomos con relación a su posición promedio (Yuan, 2005). Generalmente estos valores son usados para estudiar el esqueleto proteico (Maguid, 2006). En la representación de una molécula el espectro de colores es indicador de esta característica, donde un color azul representa la mayor rigidez y un color rojo la mayor flexibilidad.

El análisis realizado también incluye formación de estructuras secundarias a lo largo de la simulación de DM, formación de puentes de hidrógeno y distancia entre los iones Ca^{2+} y aminoácidos específicos.

5.0 Resultados

5.1 Bioinformática

A continuación, se presentan los resultados obtenidos con los métodos bioinformáticos descritos anteriormente.

5.1.1 Predicción de fosforilaciones

En la tabla 2 se muestran los aminoácidos seleccionados por el programa Netphos para estar fosforilados.

Tabla 2. Predicción de Fosforilaciones en BSP

Aminoácido	Posición	Aminoácido	Posición	Aminoácido	Posición
S	2	S	133	T	223
S	15	S	149	T	230
S	47	T	163	T	231
S	48	S	164	S	232
S	50	T	165	T	238
S	51	S	167	T	239
S	58	T	168	T	240
S	59	S	176	Y	247
T	68	S	177	Y	249
S	69	S	189	T	250
S	84	T	191	Y	255
T	89	T	201	Y	259
S	92	T	208	Y	262
T	95	T	211	S	264
Y	108	T	212	Y	281
T	124	S	214	Y	283

5.1.2 Predicción de IDR's

En la tabla 3 se muestran los resultados obtenidos con los programas PONDR y PRDOS.

Tabla 3. Predicción de IDR's en BSP humana

Parámetro	PONDR	PrDOS
Número total de residuos desordenados	183	201
Porcentaje de desorden general	60.80%	66.78%
Región desordenada más grande	106 aminoácidos	104 aminoácidos
Segmento con desorden #1	27-32	1-17
Segmento con desorden #2	46-111	43-95
Segmento con desorden #3	129-234	119-222
Segmento con desorden #4	291-295	228-229
Segmento con desorden #5	-----	236-244
Segmento con desorden #6	-----	249-273
Segmento con desorden #7	-----	290-301

5.1.3 Predicción de estructura secundaria

En la tabla 4 se muestran los resultados obtenidos con los diferentes programas. Se hace énfasis en tres posibles estructuras, mostrando el número de residuos presentes en cada estructura y el contenido de la estructura en porcentaje en relación con los aminoácidos totales de la proteína.

Tabla 4. Estructuras secundarias predichas para la BSP humana

Estructura Secundaria	Jpred		PSIPRED		CFSSP	
	(posiciones #residuos-%global)		(posiciones #residuos-%global)		(posiciones #residuos-%global)	
Hélice α	5 - 9	5 - 1.66	31-35 60-67 86-88 111-116 134-153 276-277	41 - 13.62	1-20 39-44 60-66 73-86 111-122 136-159 183-189 241-246 260-265	113- 37.54
Hoja β	190-193 199-200 209-210 246-250	13 - 4.32	198-201 210-213 241-249	17 - 5.65	27-36 89-97 106-110 208-211 222-230 235-240	43 -14.28
Giros aleatorios	1-4 10-189 194-198 201-208 211-245 251-301	283 – 94.02	1-30 36-59 68-85 89-110 117-133 154-275 278-301	243 - 80.73	21-26 37-38 45-59 67-72 87-88 98-105 123-135 160-182 190-207 212-221 231-234 247-260 266-301	145 – 48.17

5.1.4 Alineamiento múltiple de secuencias (MSA)

En las figuras 11-13, se muestra el MSA realizado para 26 especies de mamíferos, las secuencias se pueden consultar en el Anexo 2. Los aminoácidos en color rojo son los seleccionados por Netphos para presentar fosforilaciones (ver Tabla 2). También se muestran las puntuaciones de conservación por el método AACon (Valdar, 2002)

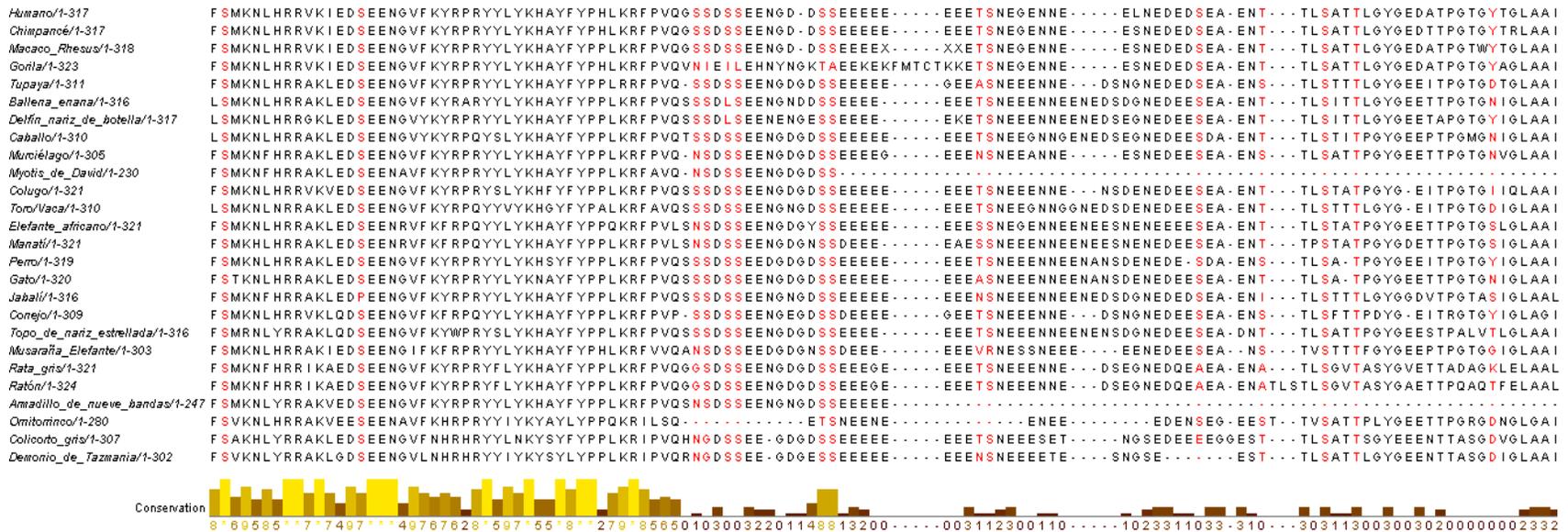


Figura 11. MSA de BSP residuos 1-114

Humano/1-317 QLPKKAAGDITNKATKEKESDE-EEEEEEENENESEAEVDENEQGINGTST--NSTEAENGNSSGGDNG-E...EGEEEEVTGANAEGETTETGRQGGKTSKTTTSPNGGFEPPTPPQ-VYR¹¹⁵TTSPFF
Chimpancé/1-317 HLPKKAAGDITNKATKEKESDE-EEEEEEENENESEAEVDENEQGINGTST--NSTEAENGNSSGGDNG-E...EGEEEEVTGANAEGETTETGQQGGKSSKTTTSPNGGFEPPTPPQ-VYR¹¹⁵TTSPFF
Macaco_Rhesus/1-318 QLPKKAAGDVTNKATKKEESDE-EEEEEEENENESEAEVDENEQGINGTST--NSTEAENGNSSGGDNGEE...EGEEEEVTGASAEGETTATGGQGGKSSKTTTSPNGGFEPPTPPQ-VYR¹¹⁵TTSLPF
Gorila/1-323 QLPKKAAGDITNKATKEKESDE-EEEEEEENENESEAEVDENEQGINGTST--NSTEAENGNSSGGDNG-E...EGEEEEVTGANAEGETTETGGQGGKTSKTTTSPNGGFEPPTPPQ-VYR¹¹⁵TTSPFF
Tupaya/1-311 QLPKKAAGDKRNKATKQKESDE-EEEEEEENENESEAEVDENEQGINGTST--NSTEVDNNGSSGGDNGEE...EGEEEEVTGAVAEGETTVTRGQDNGSSKTTTSSNGLVPTPPQ-IY¹¹⁵GT...
Ballena_enana/1-316 QLPKKAAGDIGKKVTKEESD...EEEEEEEN--ENEAEADDNEQGINGTST--NGTEIDNNGHSSGGDNGEE...EGEE-GVTEANAEGTTVAGQDNGSSKATTSPNGGFEPPTPPQ-IDG¹¹⁵TTPPFF
Delphin_nariz_de_botella/1-317 QLPKKAAGDIGKAATEEESD...EEGGGEEEN--ENEAEVDNNEQGVN¹¹⁵TSA--NSTEIDNNGHSSGGDNGEE...EGEE-DVTEANAEGTTVAGQDNGSSKATTSPNGGFEPPTPPQ-IY¹¹⁵GTTPPPS
Caballo/1-310 QLPKKAAGDKGNKGTKEESD...EEEEEEEDENESEAEVDENEQGINGTSS--NSTEAENGNSSGGDNGEE...EGEE...EANAEGTAVPAEQDNGSSKTTTSPNGGFAPPTPPQEVYR¹¹⁵...
Murciélago/1-305 QLPKKAAGDVGKKATKKEESE-EEEEEEENENESEAEVDENEQGINGTST--NSTDVENGNSSGGDNGEE-EEEEEE...-NAEGTTVAGEQD...-N¹¹⁵GGFEPPTPPQEVYGTTPPPQ
Myotis_de_David/1-230 ...EEEEEEENENESEAEVDENEQGVN¹¹⁵TST--NSTDAENGNSSGGDNGEEEGEEEE...-NAEGTTVAGEQD...-N¹¹⁵GGSEPTTPPQAYGTTPPPQ
Colugo/1-321 HLPKKAAGDIGKKAQKQESDE-EEEEEEEDENESEAEVDENEQGINGTST--NSTEVDNNGHSSGGDNGEE...DGE¹¹⁵GVTDANGEGTAPGEWDNGSSSETTSPNGGFQPTTPPLEGYGTTPPPF
Toro/Vaca/1-310 WLPKKAAGTGGKATKKEESD...EEEEEEENENESEAEVDENEQGINGTSS--NSTEVDNNGHSSGGDNGEE...DGE¹¹⁵SVTEANTEGITVAG...-ETTSPNGGFKPTPHQEVYGTTPPPF
Elefante_africano/1-321 QLPKKAAGHTGNKVTKTEESD...EEEEEEENENESEAEVDENEQGINGTST--NSTDTENGNSSGGDNGEE...EGEEEEVTANAEEATVTGEQNGGLETTASPNAGFQPTAPPQDIY¹¹⁵GTTPPF
Manatí/1-321 QLPKKAAGDIGKKVTKKEESD...EEEEEEENENESEAEVDENEQGINGTST--NSTDAENGNSSGGDNGEE...EGEEEEVTANAETATMTGEQDNGGVETTTSPNDGFEPPTPPQDIY¹¹⁵GTASPF
Perro/1-319 QLPKKAAGDIRHKATKEESD...EEEEEEEDENESEAEVDENEQGINSTSS--NSTEAENGNSSAGDNG...EGEEEEVTAEHSEGTTEAGKQNNSSKTTLSPDGGFEPPTPPPELYGTTRPS
Gato/1-320 QLPKKAAGDTKKATREESD...EEEEEEENENESEAEVDENEQGINSTSTNSTEVENGNSSDNGTGE...EGEEEEVTENAEGTTMAGRQDNGSTTTSPNGGFQPTTPPPEFYGTTRPS
Jabalí/1-316 QLPKKAAGDIGKKSACEESD...EEEEEEENENESEAEVDENEQGTNGTST--NSTEVDNNGHSSGGDNGE...EGDQ¹¹⁵SVTEA--Q¹¹⁵TTVAGEQDNGGAKTTTSPNGGLEPTPPQDIS¹¹⁵TTLPPS
Conejo/1-309 KLPKKAAGAIENKATEEESD...EEEEEEENENESEAEVDENEQGINGTSS--NSTEAENGE...-AGE¹¹⁵ESVTATDAAGTATRGEWISRAE¹¹⁵TTAPSGGFEPPTPPPEAYGTTPPPF
Topo_de_nariz_estrellada/1-316 QLPKKAAGPTGNKSTKKEESD...EEEEEEENENESEAEVDENEQGINGTST--NSTEVDNNGHSSGGDNGE...EGEE-SATESNAEATTVAGRQDKA¹¹⁵ITKTTSPNGGFEPPTPPPEYGTTPPPL
Musaraña_Elefante/1-303 QLPKKAAGDAGSKATKQESD...EEEEEEENENESEAEVDENEQGINGTST--NSTDVENGNSSGGDNGEE...EGEEENVTKASVEGTTAPAEQ...-ETT¹¹⁵YANGRFELTTT...-F
Rata_gris/1-321 QLPKKAAGDAEGKAPKMKESDEEEEEEEENENESEAEVDENEQVVNGTST--NSTEVDGNGSSGGDNGE...EAE¹¹⁵ESVTEAGAE¹¹⁵TTAGARELTSYGT¹¹⁵TTAVLLNGFQQTTPPPEAYGTTPSPA
Ratón/1-324 QLPKKAAGDAESRAPKVKESDEEEEEEEENENESEAEVDENELAVNGTST--NSTEVDGNGSSGGDNGE-A-EAE¹¹⁵ESVTEAGAE¹¹⁵TT-GGRELTSVGTQTAVLLNGFQQTTPPPEAYGTTPSPI
Amadillo_de_nueve_bandas/1-247 ...-AGNIG...SKEEESD...EEEEEEENENESEAEVDENEQGVN¹¹⁵TST--DSTDVENGNSSGGDNG...EGE...-TEDNPEGTTTIGQDNGRFKTTASPNAGFEPPTPSHKIYR¹¹⁵TTPSY
Omitomino/1-280 KVPKKA--ETRKSTKKEESDEEEEEEEENENESEAEVDENEQGINGTST--NSTE-D¹¹⁵SGN--D¹¹⁵GGENGGE-AAEEEEESVTEANQE...-ENG¹¹⁵QETTT¹¹⁵FVDSVQGATTPSNGAYE¹¹⁵VTTP
Colicorbo_gris/1-307 QIPKK-GEDSKNNTKKEESDEEEEEEEENENESEAEVDKNKEVVNGTDH...-D¹¹⁵GGNGTS¹¹⁵GKEDGEE...EEVEN¹¹⁵GTSSVD...-KDN¹¹⁵STELATSPNSRYESTPLEGYE¹¹⁵VTSPY
Demonio_de_Tazmania/1-302 QNPKK-EEKNNQVTKKEESD...EEEEEEENENESEAEVDENKGGNS¹¹⁵NH--ENAEGDHGNNGGKEEGEE...EEVEN¹¹⁵GTTEATEN...-EAS¹¹⁵STV¹¹⁵TSLNG-YEFT¹¹⁵LLKGD¹¹⁵FEV¹¹⁵TSPPY

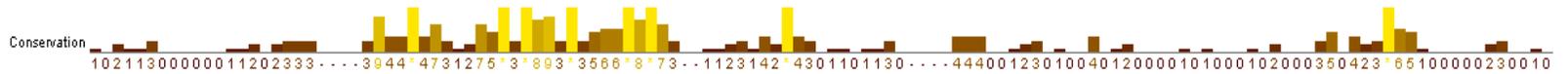


Figura 12 MSA de BSP residuos 115-235



Figura 13. MSA de BSP residuos 236-301

5.1.5 Selección de péptidos

Para la selección de los péptidos se consideraron los siguientes puntos: fuerzas electrostáticas (fosforilaciones y aminoácidos con carga), regiones desordenadas, estructuras secundarias y zonas conservadas en el MSA. Además de la información existente acerca de la BSP.

En la tabla 5 se muestran las regiones de la BSP, consideradas para la construcción de péptidos:

Tabla 5. Secuencias seleccionadas para la construcción de péptidos para simulaciones de DM

Secuencia	Posiciones	Péptido
KNLHRRVKIEDSEENGVFKY	4-23	1
RPRYYLYKHAYFYPHLKRFP	24-43	2
RPRYYLYKHAYFYPHLKRFPVQGSSDSSEENGDDSSSEEEE	24-63	3
VQGSSDSSEENGDDSSSEEEEEEEETSNEGENNEELNEDED	44-83	4
NKATKEKESDEEEEEEEEGN	125-144	5
IYSENGEPRGDNYRAYEDEYSYFKGQGYDGYDGQNYHH	261-300	6

Las secuencias mostradas en la tabla 5, representan los distintos péptidos a usar en las simulaciones de DM. En la tabla 6 se muestran modificaciones realizadas a estos péptidos, las cuales serán referidas como versiones de los péptidos:

Tabla 6. Modificaciones realizadas a los péptidos seleccionados para su uso en simulaciones de DM

Secuencia	Carga total	Nombre
KNLHRRVKIEDSEENGVFKY	-1	P1
KNLHRRVKIEDSEENGVFKY	-1	P1Ca
KNLHRRVKIEDSpEENGVFKY	-3	P1F
KNLHRRVKIEDSpEENGVFKY	-3	P1FCa

RPRYYLYKHAYFYPHLKRFP	+5	P2
RPRYYLYKHAYFYPHLKRFP	+5	P2Ca
RPRYYLYKHAYFYPHLKRFPVQGSSDSSEENGDDSSSEEEE	-4	P3
RPRYYLYKHAYFYPHLKRFPVQGSSDSSEENGDDSSSEEEE	-4	P3Ca
RPRYYLYKHAYFYPHLKRFPVQGS _p SpD _p SpSpEENGDD _p SpE _p EEE	-16	P3F
RPRYYLYKHAYFYPHLKRFPVQGS _p SpD _p SpSpEENGDD _p SpE _p EEE	-16	P3FCa
VQGSSDSSEENGDDSSSEEEEEEEETSNEGENNEELNEDED	-21	P4
VQGSSDSSEENGDDSSSEEEEEEEETSNEGENNEELNEDED	-21	P4Ca
VQGS _p SpD _p SpSpEENGDD _p SpE _p EEEEEEET _p SpNEGENNEELNEDED	-37	P4F
VQGS _p SpD _p SpSpEENGDD _p SpE _p EEEEEEET _p SpNEGENNEELNEDED	-37	P4FCa
NKATKEKESDEEEEEEEEGN	-11	P5
NKATKEKESDEEEEEEEEGN	-11	P5Ca
NKATKEKES _p DEEEEEEEEGN	-13	P5F
NKATKEKES _p DEEEEEEEEGN	-13	P5FCa
IYSENGEPRGDNYRAYEDEYSYFKGQGYDGYDGQNYHH	-5	P6
IYSENGEPRGDNYRAYEDEYSYFKGQGYDGYDGQNYHH	-5	P6Ca
I _p ES _p ENGEPRGDNYRAYEDEY _p SY _p FKGQGYDGYDGQNYHH	-13	P6F
I _p ES _p ENGEPRGDNYRAYEDEY _p SY _p FKGQGYDGYDGQNYHH	-13	P6FCa

Los aminoácidos con el subíndice “p” son los señalados por NetPhos para ser fosforilados (Tabla 2). En la segunda columna se muestra la carga total del péptido, obtenida de los archivos de topología. La tercera columna indica los códigos asignados a cada péptido y a sus respectivas modificaciones, donde el número representa el número del péptido, F la presencia de fosforilaciones, mientras que el sufijo “Ca” indica la adición de iones Calcio²⁺ a una concentración de 0.1M en la caja de simulación. Los péptidos en la tabla se seleccionaron cumpliendo una estandarización en los protocolos utilizados por el grupo de trabajo (Villarreal-Ramírez, 2014; Boskey, 2016)

En la sección 6.0 se muestra la discusión concerniente a la selección de péptidos. Los modelos de los péptidos seleccionados se construyeron con la paquetería “The PyMOL Molecular Graphics System, Version 1.2r3pre (Schrödinger)” y las fosforilaciones se añadieron a los modelos con la paquetería de YASARA (Krieger,

2014). A los extremos de los se péptidos se agregó un *capping*: un grupo acetilo y un grupo amino en el carboxilo y amino terminal, respectivamente.

5.2 Simulaciones de DM

A continuación, se muestran los resultados obtenidos para los diferentes péptidos estudiados, se analizarán sus características estructurales, efecto de las fosforilaciones e interacciones con iones Ca^{2+} .

Se realizaron simulaciones de DM de 30ns, donde los primeros 5ns se consideraron como el tiempo de estabilización, determinado por varias simulaciones de prueba con distintos péptidos. A excepción de los valores iniciales mostrados para las trayectorias, puentes de hidrógeno y para el computo de distancias. Para cada péptido se realizaron simulaciones de DM, en su versión fosforilada, sin fosforilar y en ausencia y presencia de iones Ca^{2+} a una concentración de 0.1M. En la tabla 6 se presenta el código de nombres utilizado para los péptidos. Además, los resultados gráficos siguen un código de color:

- Rojo: Péptido no fosforilado.
- Negro: Péptido fosforilado.
- Azul: Péptido no fosforilado en presencia de iones Ca^{2+} .
- Verde: Péptido fosforilado en presencia de iones Ca^{2+} .

En las imágenes de los péptidos, los átomos de las cadenas laterales se representan por los siguientes colores:

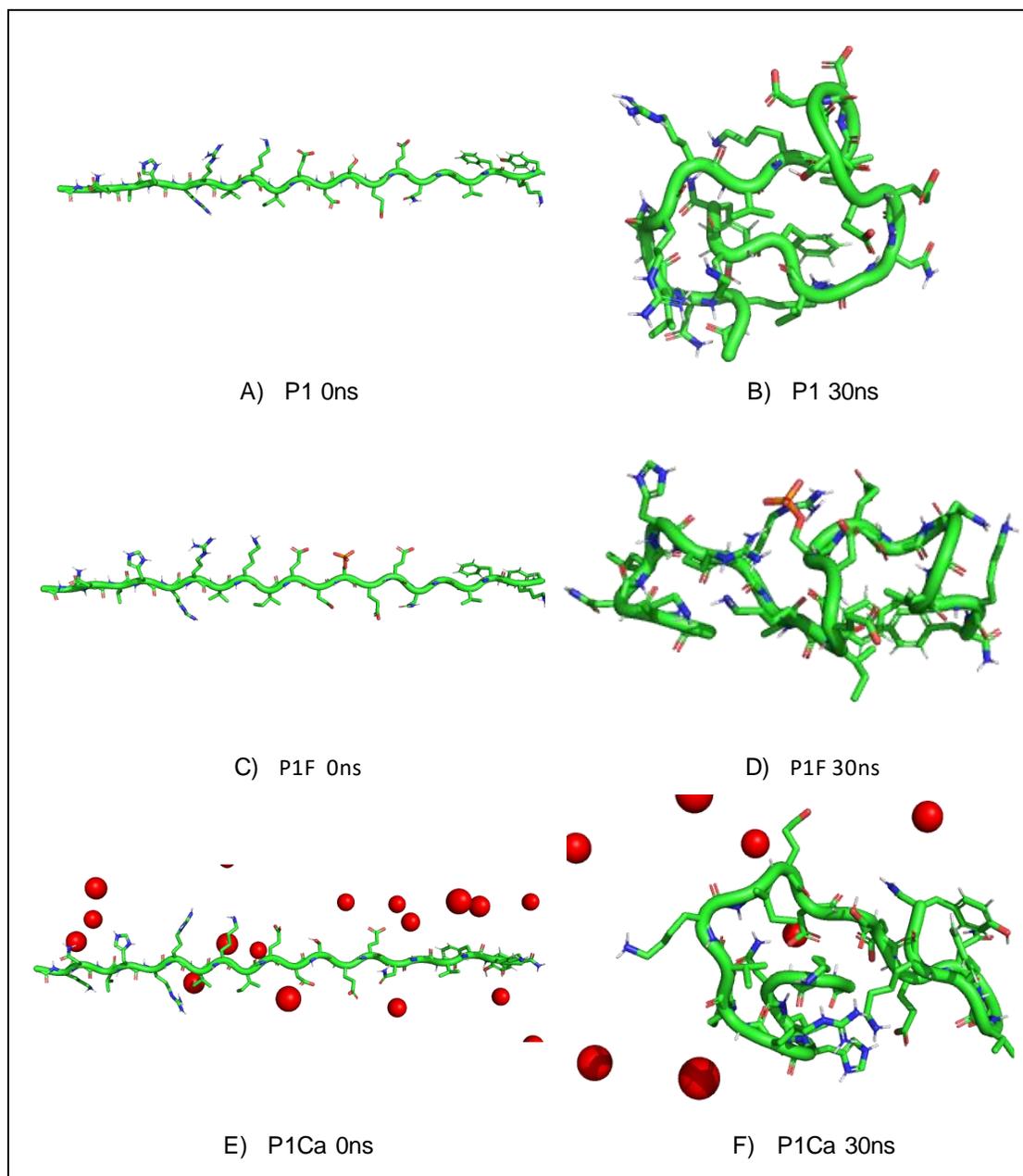
- Rojo: Oxígeno.
- Naranja: Fosforo.
- Azul: Nitrógeno.
- Blanco: Hidrógeno.
- Verde: Carbono.

El esqueleto de la cadena principal del péptido se encuentra representado gráficamente por una cinta de color verde.

5.2.1 Péptido 1

Se muestran imágenes que corresponden al estado estructural correspondiente al inicio y al final de la dinámica, las trayectorias fueron visualizadas en PyMOL para posteriormente exportar el *frame* deseado.

En la figura 14 (B,D,F y H) se observa como la estructura de loops es frecuente en todas las versiones del péptido 1 (P1, P1F, P1Ca y P1FCa, tabla 6)



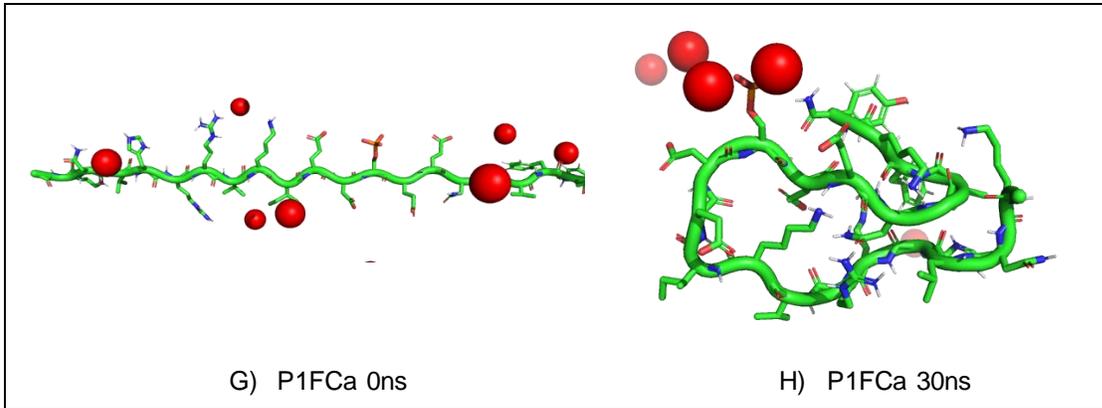


Figura 14. Estados estructurales de las distintas versiones del péptido 1 (Tabla 6). La columna izquierda muestra el estado inicial, mientras que la columna derecha muestra las estructuras adquiridas al transcurrir los 30ns de simulación.

Radio de giro (Rg)

El grado de compactación del péptido 1 en todas las versiones se representa en la figura 15:

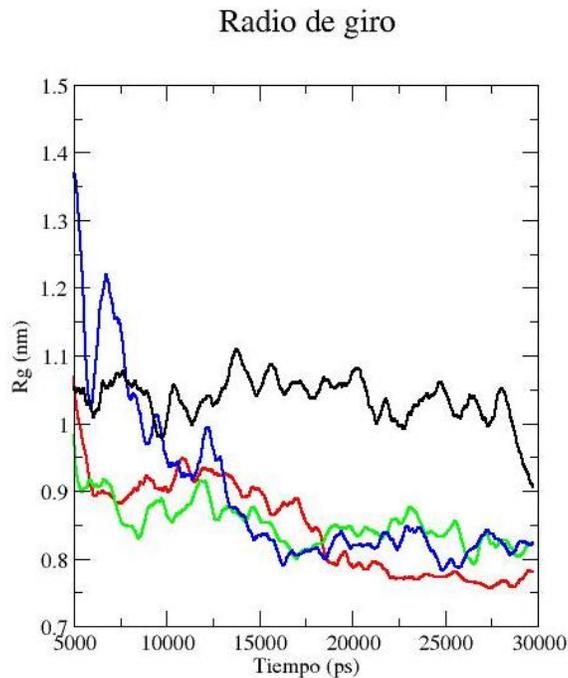


Figura 15. Radio de giro péptido 1.

El código de colores para las distintas versiones del péptido es P1 - rojo, P1F - negro, P1Ca – azul, P1FCa – verde.

En la figura 15 se observa a partir de los 10ns como el péptido P1F tiene el mayor Rg respecto a las otras versiones del péptido. P1F es el péptido con menor compactación. Sin embargo, P1FCa tiene un grado de compactación mayor. El lector deberá recordar que P1FCa es idéntico a P1F, la diferencia estriba en que en la caja de simulación de P1F no hay presencia de iones Ca^{2+} . P1 es el péptido con mayor compactación, con un Rg promedio de 8.46 Å. Sin embargo, P1Ca tiene una diferencia de Rg de 0.47 Å, debido a la adición de los iones Ca^{2+} .

Para comprender el comportamiento y la movilidad de cada residuo de aminoácido durante las simulaciones de DM se estudió el RMSF y el valor del factor B.

RMSF y Factor B

En la figura 16, se muestran los resultados del análisis del RMSF de los péptidos P1 (rojo), P1F (negro), P1Ca (azul) y P1FCa (verde). La fosforilación favorece la flexibilidad del péptido P1F, con respecto a P1. P1F presenta una mayor movilidad en los aminoácidos K-4, N-5, L-6, V-10, K-11, I-12, E-13, D-14, S-15, V-20, F-21 y K-22. Donde los residuos de aminoácido K-4 y E-13 tienen la mayor movilidad, 6.153 Å y 6.075 Å, respectivamente, en P1F. Mientras, los aminoácidos K-4 y E-13 en la versión de P1 presentan una movilidad de 58.18 Å y 38.58 Å, respectivamente. El péptido P1FCa muestra un aumento de flexibilidad para 16 aminoácidos en comparación al péptido P1. El péptido P1FCA muestra ligeros aumentos y disminuciones de flexibilidad, en donde la región con mayores cambios está representada por la disminución de flexibilidad en los aminoácidos S-15, E-16, E-17, N-18 cuya diferencia a sus aminoácidos homólogos en el péptido P1 es de 0.738 Å, 2.517 Å, 1.847 Å y 1.931 Å, respectivamente.

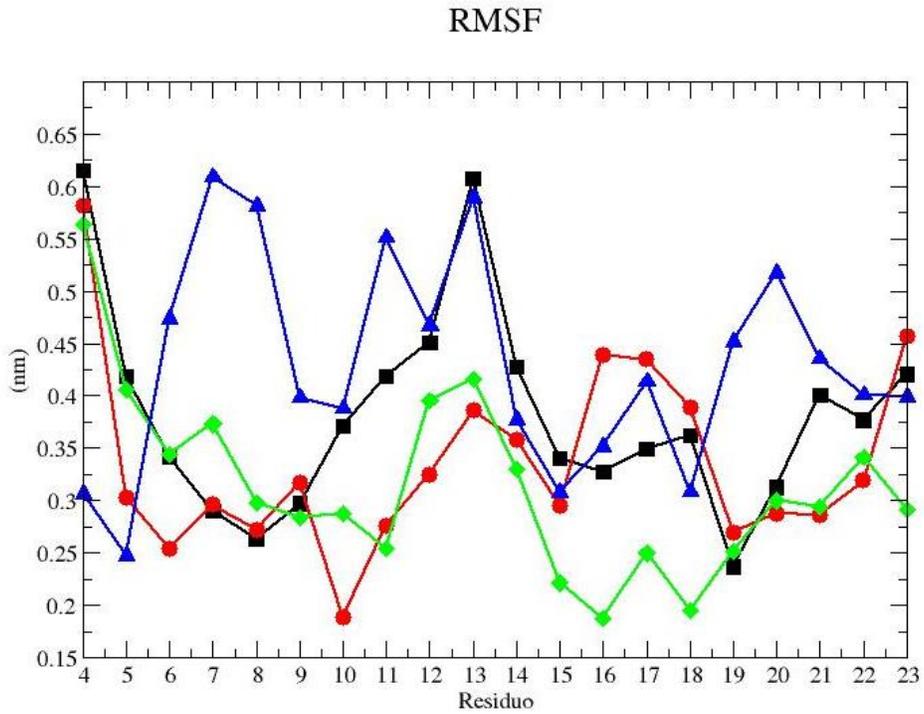
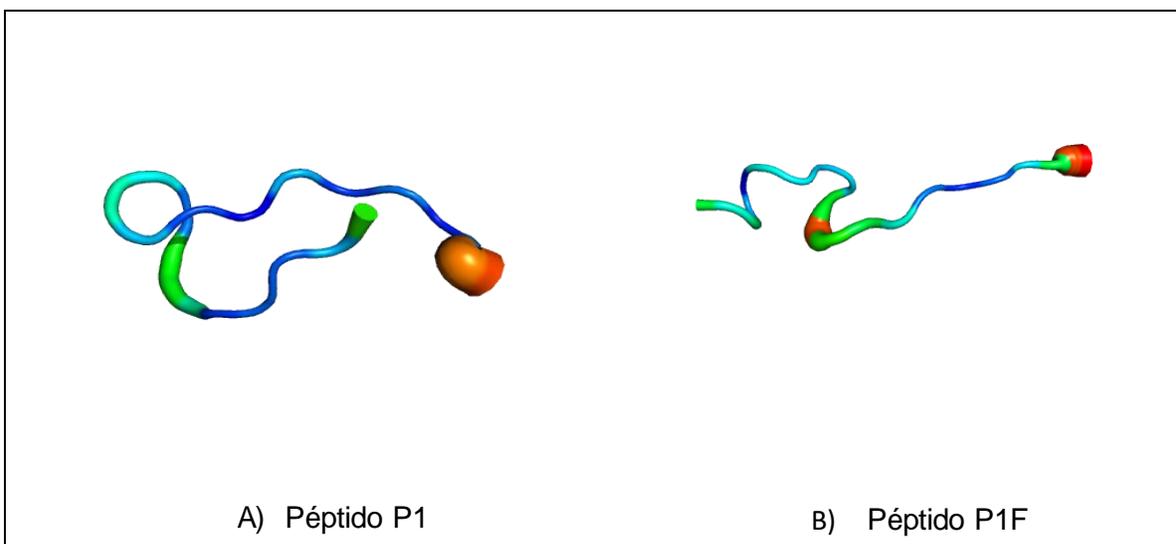


Figura 16. RMSF péptido 1

Se sigue la convención de colores especificada al inicio.

El RMSF junto al factor B nos sirven para evaluar en conjunto la flexibilidad de los péptidos, a continuación, en la figura 17 (A-D), se observan los factores B y la estructura promedio para los péptidos P1, P1F, P1Ca y P1FCa:



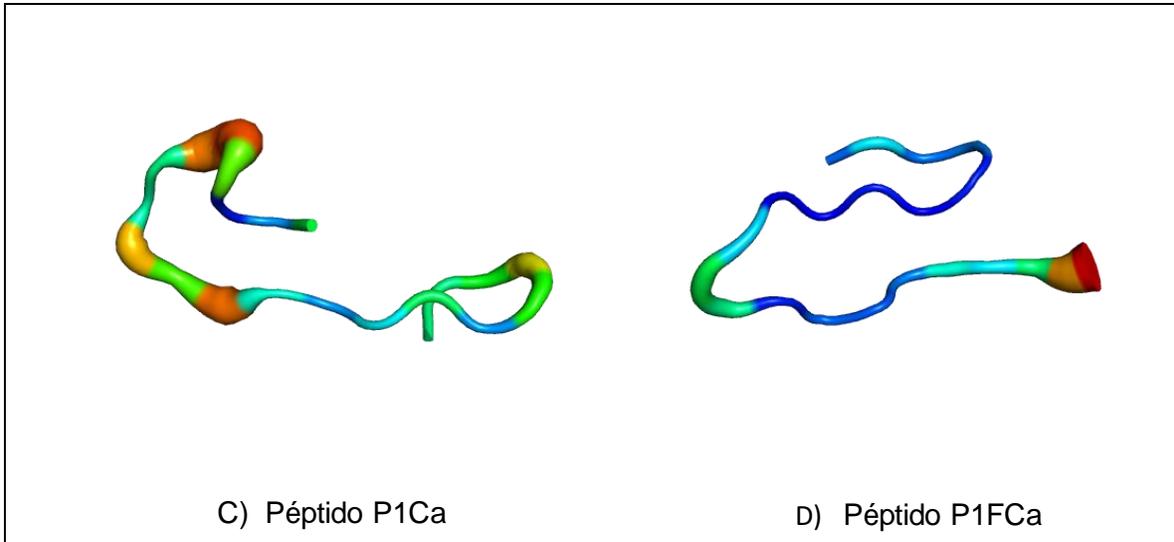


Figura 17. Estructura promedio de las diferentes versiones del péptido 1 y factor B representado por un espectro de colores azul-rojo

El espectro de colores mostrado en la figura 17 A-D, guardan relación con el grado de rigidez o flexibilidad presentada por cada residuo de aminoácido, donde un color rojo indica la mayor movilidad y el azul la mayor rigidez. El grosor mostrado en el “modelo de cinta” es proporcional a los valores de RMSF, mediante los cuales se obtuvo el factor B (Carluccio, 2013). Los factores B para las 4 versiones del péptido 1 se calcularon bajo la misma escala, por lo que la diferencia de colores en los distintos péptidos (Figura 17 A-D) es un indicativo directo de cambios de flexibilidad existentes entre ellos.

En las figuras 17 A y B se observan los cambios de flexibilidad entre el P1 y el P1F producto de la adición de un grupo fosfato en la S-15. De manera general, la zona anterior a la Sp-15 del péptido P1F, correspondiente a los aminoácidos RVKIEDSp presenta un aumento en la flexibilidad, identificada por una coloración verde-naranja, en comparación con los colores azules en el P1, mientras que la zona posterior a la Sp-15, correspondiente a los aminoácidos EEN, presenta un decremento de flexibilidad exhibido por color azul en el P1F y de color verde en el péptido P1. En las figuras 17 A y C, correspondientes a los péptidos P1 y P1Ca, respectivamente, se observa un aumento de flexibilidad en distintos aminoácidos a lo largo de la secuencia del P1Ca, lo que se identifica por la presencia de colores

amarillos, naranjas y verdes que no se observan en el péptido P1 (Figura 17 A). El péptido P1FCa (Figura 17 D), presenta una disminución de flexibilidad en la región IEDSEEN, la cual está representada por la presencia de tonalidades azul en el modelo de cinta, en comparación con los colores verdes presentes en el péptido P1 (Figura 17 A).

Puentes de Hidrogeno (PH)

La formación de PH en las proteínas contribuye a la formación de la estructura secundaria (Pace, 2014). El análisis de los PH se realizó con el programa HBonanza (Durrant, 2011) y la herramienta *hbond* de GROMACS. Los resultados se encuentran resumidos en la tabla 7. Únicamente se mostrarán aquellos puentes con una aparición de 25% del tiempo total de la simulación.

Tabla 7. PH formados en las distintas versiones del péptido 1 (P1,P1F, P1Ca y P1FCa). La numeración de los aminoácidos corresponde a la secuencia de la BSP.

Péptido	Puentes formados (Donador – Aceptor)	Aparición del puente durante la dinámica (%)
Péptido sin fosforilar (P1)	Ser15 – Lys11	46.3
	Arg9 – Lys11	34.7
	Arg9 – Lys11	30.5
	Phe21 – Lys4	28.1
	Lys18 – Arg8	27.2
	Val10 – Val20	25.6
Péptido fosforilado (P1F)	Arg9 – Sep15	71.0
	Arg9 – Sep15	66.9
	Arg8 – Sep15	60.3
	Arg8 – Sep15	56.2
	Asp14 – Ile13	53.7
	Arg8 – Lys11	44.6
	Gly19 – Lys22	38.8
	Lys22 – Glu17	36.3
	Arg8 – Lys11	34.7
	Ile12 – Asp14	33.0

Péptido sin fosforilar con Ca²⁺ (P1Ca)	Tyr23 – Glu16	54.5
	Asn18 – Phe21	52.0
	Arg8 – Lys4	48.7
	His7 – Glu17	40.5
	Asn5 – Asp14	37.2
	His7 – Glu17	31.4
	Glu16 – Tyr22	29.8
	Phe21 – Asn14	27.2
	Asn5 – Asp14	25.6
Péptido fosforilado con Ca²⁺ (P1FCa)	Tyr23 – Glu16	54.5
	Phe21 – Asn18	49.6
	Asn18 – Phe21	47.1
	Arg9 – Lys11	46.3
	Glu16 – Sep15	44.6
	Arg9 – Lys11	41.3
	Lys11 – Glu16	35.5
	Asn18 -GLu16	26.4

De la tabla 7 se observa que la Sep-15 del P1F participa en la formación de cuatro puentes de hidrógeno y con una presencia de más del 50% durante las simulaciones de DM. Sin embargo, Sep15 en P1FCa sólo participa en la formación de 1 PH. La Ser-15 en P1 participa en la formación de 1 PH y en P1Ca no se ve involucrada en la formación de ningún PH.

Para evaluar la movilidad de los iones Ca²⁺ durante las simulaciones de DM, se estudiaron los factores B. En este trabajo, se consideraron a los iones Ca²⁺, con una baja movilidad, con una alta probabilidad de tener interacciones con los péptidos, dichos iones se observan en color azul en la figura 18. Los iones Ca²⁺ de color rojo en la figura 18 son aquellos con mayor movilidad, por lo que se asume que tienen interacciones transitorias o nulas con el péptido. La figura 19, muestra a manera de ejemplo, los factores B para los iones Ca²⁺ del péptido P1FCa y se excluyó la presencia de éste con fines didácticos. De ahora en adelante, no se vuelven a mostrar los factores B de los iones.

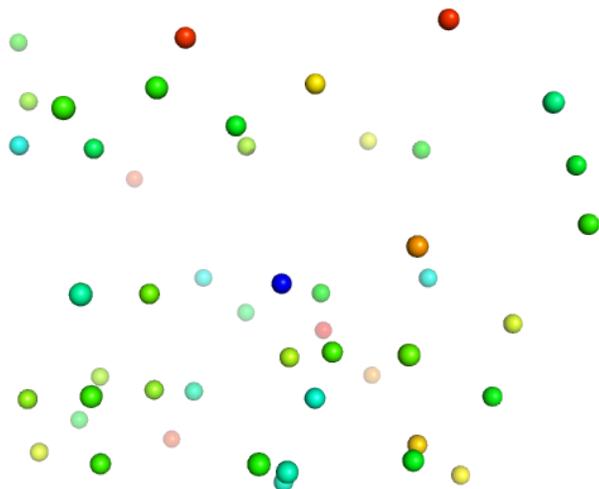


Figura 18. Factor B de iones Ca^{2+} de la simulación de DM del péptido P1FCa.

Obtenidos para comprobar el grado de asociación con el péptido, se revisó la trayectoria del péptido y el comportamiento de iones específicos durante la simulación.

Se revisó la trayectoria de los péptidos y el comportamiento de aquellos iones identificados con posibles interacciones con cada péptido.

Como era esperado, los aminoácidos con una carga negativa estuvieron asociados a los iones calcio de manera transitoria y/o formando un complejo. Los iones Ca^{2+} asociados a manera de complejo se seleccionaron para estudiar la distancia entre el centro de masas (COM) del ion y el centro de masas del residuo. Mientras los iones Ca^{2+} con interacciones momentáneas no son considerados. Los resultados se muestran en la tabla 8.

Tabla 8. Distancia entre residuos y iones Ca^{2+} formadores de complejo

Péptido	Ca^{2+} en interacción	Residuo con el que interactúa	Distancia inicial COM residuo – COM Ca^{2+} (Å)	Distancia mínima COM residuo – COM Ca^{2+} (Å)
Péptido 1 fosforilado	5501	Sep13	52.23	3.16
	5480	Sep13	36.22	9.55

Estructura secundaria

Como parte de los análisis empleados se utilizó la herramienta DSSP de GROMACS para determinar el contenido de estructura secundaria en los diferentes péptidos a lo largo de las simulaciones de DM. Los resultados se muestran en la Tabla 9:

Tabla 9. Determinación de estructuras secundarias en el péptido 1

Péptido	DSSP		
	Estructura determinada		
	Hélice α	Hoja β	Giros
P1	1%	45%	54%
P1F	5%	43%	52%
P1Ca	0%	52%	48%
P1FCa	0%	41%	59%

5.2.2 Péptido 2

De la figura 19 B y D se observa que ambos péptidos tienden a formar estructuras de giros y bucles.

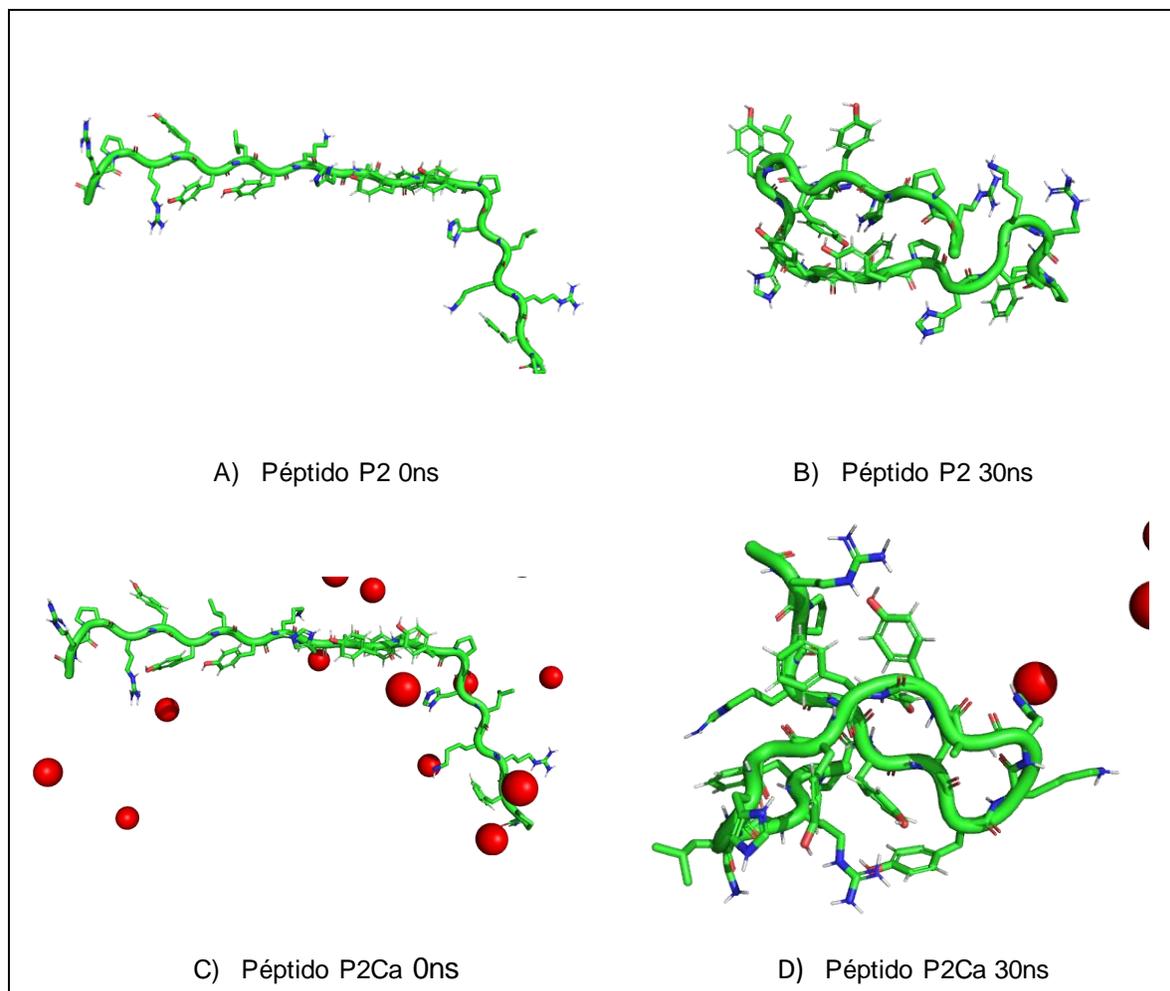


Figura 19 Estados estructurales de las distintas versiones del péptido 2 (Tabla 6)

La columna izquierda muestra el estado inicial, mientras que la columna derecha muestra las estructuras adquiridas al transcurrir los 30ns de simulación.

Radio de giro

El grado de compactación de los péptidos P2 y P2Ca se muestra en la Figura 20. Se observa que posterior a los 15ns hay un cambio en el comportamiento de los péptidos, donde el péptido P1Ca (azul) adquiere un mayor grado de compactación, con respecto al péptido P1 (rojo).

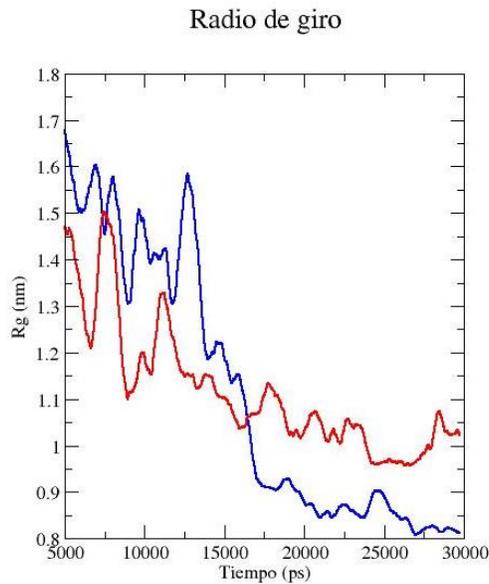


Figura 20 Radio de giro péptido 2

RMSF y Factor B

En la figura 21 se muestran los resultados del análisis del RMSF de los péptidos P2 y P2Ca. Se observa que la presencia de iones Ca^{2+} en el péptido P2Ca provoca un aumento en la movilidad en los aminoácidos con respecto al péptido P2 donde no hay presencia de estos.

De la figura 22 A y B se observa que la presencia de iones Ca^{2+} en el péptido P2Ca, confiere una mayor flexibilidad al péptido, con respecto al péptido P2, siendo coherente con la figura 21. Los cambios en la movilidad a lo largo de la cadena del péptido P2Ca (Figura 22 B) se identifican por la presencia de colores amarillos, verdes y tonalidades de azul más claras, en contraste con las predominantes tonalidades oscuras de azul, presentes en el péptido P2 (Figura 22 A).

RMSF

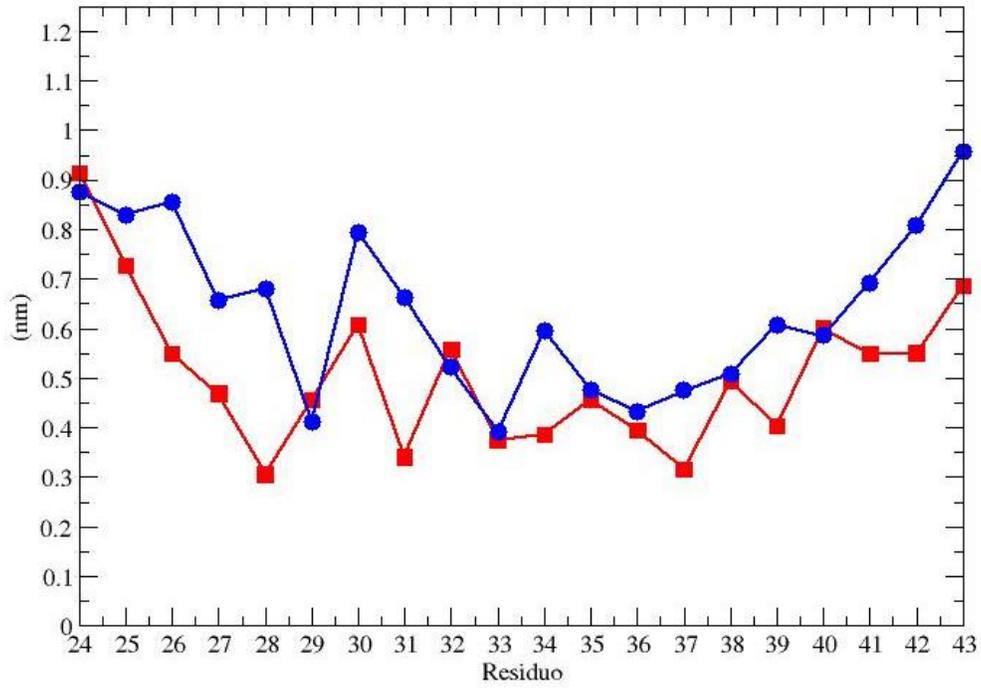


Figura 21. RMSF péptido 2

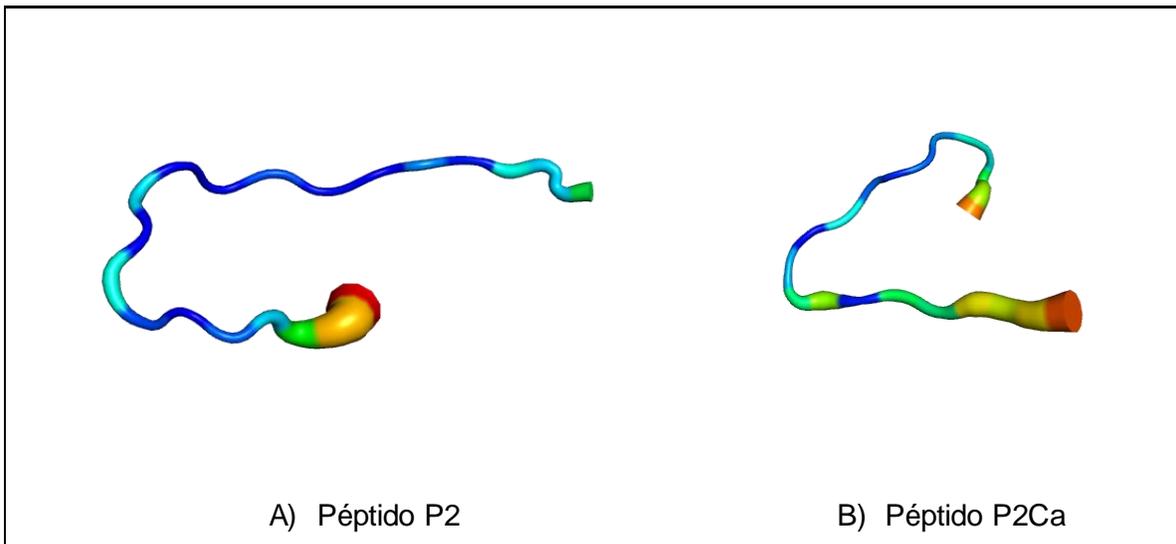


Figura 22. Estructura promedio de las diferentes versiones del péptido 2 y factor B representado por un espectro de colores azul-rojo

PH

A continuación, se muestran los PH computados para ambos péptidos (P2 y P2Ca).

Tabla 10. PH formados en las distintas versiones del péptido 2 (P2, P2Ca)

Péptido	Puentes formados (Donador – Aceptor)	Aparición del puente durante la dinámica (%)
P2	Phe13 – Tyr6	49.6
	Arg2 – Pro15	32.2
	Arg4 – Lys9	28.1
P2Ca	Tyr6 – Pro21	26.4

El péptido en ausencia de calcio, P2, presenta más puentes de hidrógeno y con una mayor duración que los formados por el péptido P2Ca.

Se buscó evaluar las interacciones con iones Ca^{2+} , como se explicó en la sección correspondiente al péptido 1, pero no se encontraron iones que se asociaran a modo de manera directa con el péptido.

Estructura secundaria

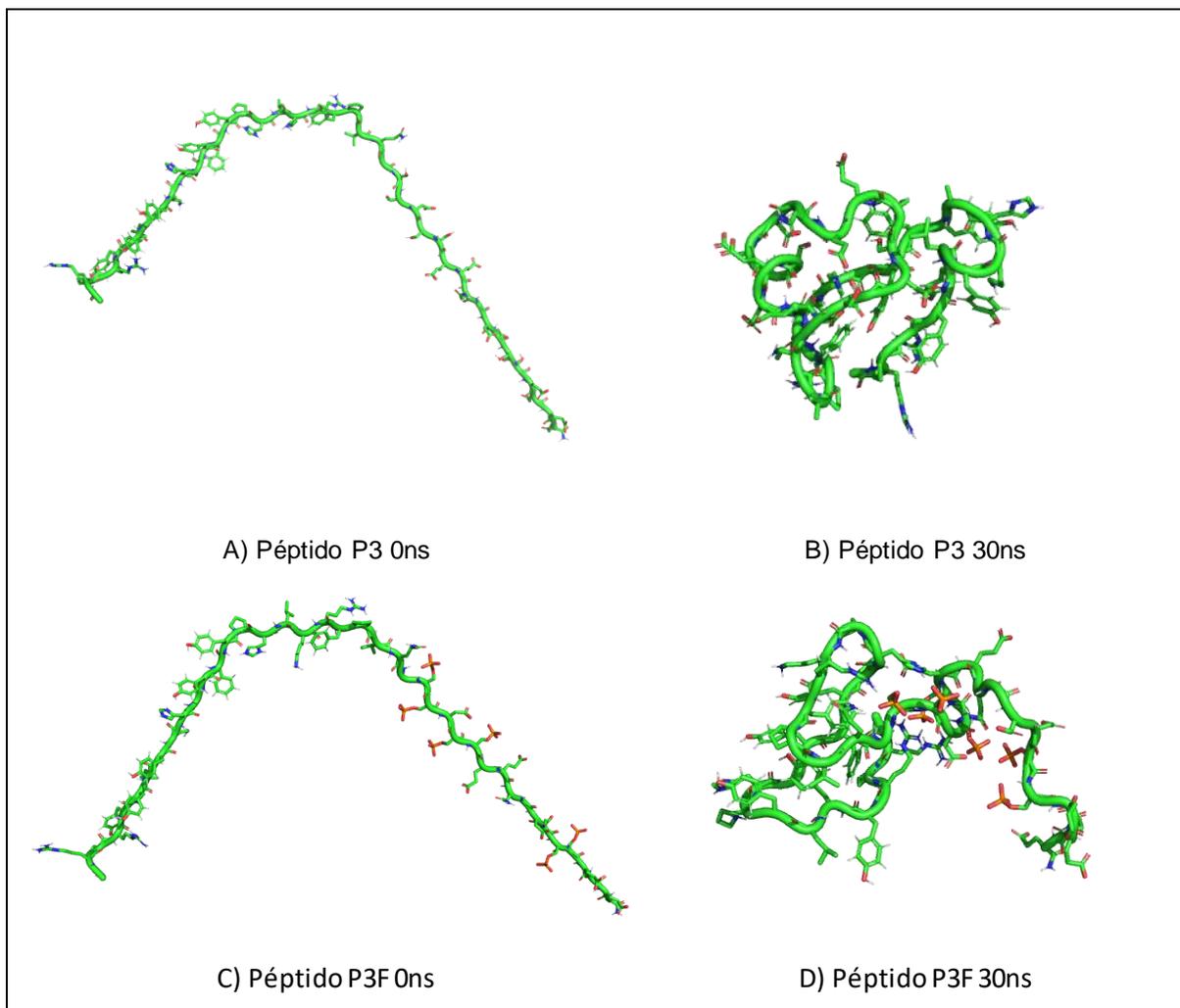
En la tabla 11 se muestran los contenidos de estructura secundaria según DSSP:

Tabla 11. Determinación de estructuras secundarias en el péptido 2

Péptido	DSSP		
	Estructura determinada		
	Hélice α	Hoja β	Giros
P2	0%	21%	79%
P2Ca	1%	27%	72%

5.2.3 Péptido 3

En primer lugar se muestran las imágenes de diferentes estados estructurales del péptido en sus distintas versiones (P3, P3F, P3Ca y P3FCa):



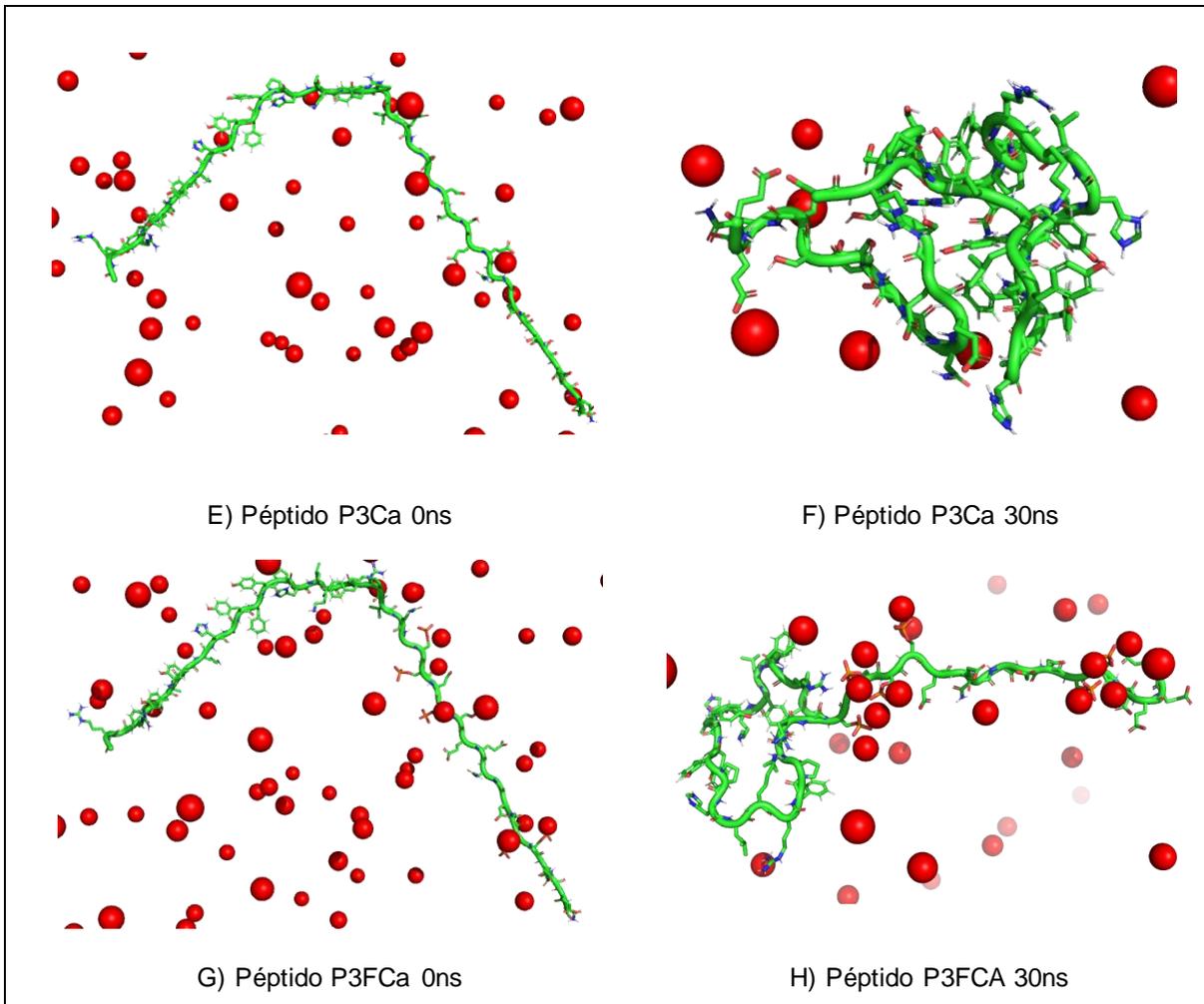


Figura 23 Estados estructurales de las distintas versiones del péptido 3 (Tabla 6)

La columna izquierda muestra el estado inicial, mientras que la columna derecha muestra las estructuras adquiridas al transcurrir los 30ns de simulación.

Se observa que los péptidos tienen a formar estructuras tipo *loop* y lo que más destaca es la conformación que adquiere el péptido P3FCa (Figura 23 H), donde la región que contiene a las serinas fosforiladas (SpSpDSpSpEENGDDSpSpEEEE), se encuentra extendida y rodeada de iones Ca^{2+} .

Radio de giro

El uso del gráfico de radio de giro nos permite tener una idea más clara de la compactación del péptido. De la figura 24, se observa que la presencia de iones Ca^{2+} aparentemente no provoca cambios mayores entre la compactación del

péptido P3Ca (azul) y P3 (rojo), teniendo un Rg promedio de 10.33A y 10.25A, respectivamente. El péptido P3F (negro) presenta un mayor Rg a lo largo de toda dinámica que el del péptido P3, lo que se interpreta como una menor compactación. El péptido P3FCa tiene la menor compactación de todos los péptidos al tener el mayor Rg a lo largo de toda la simulación de DM.

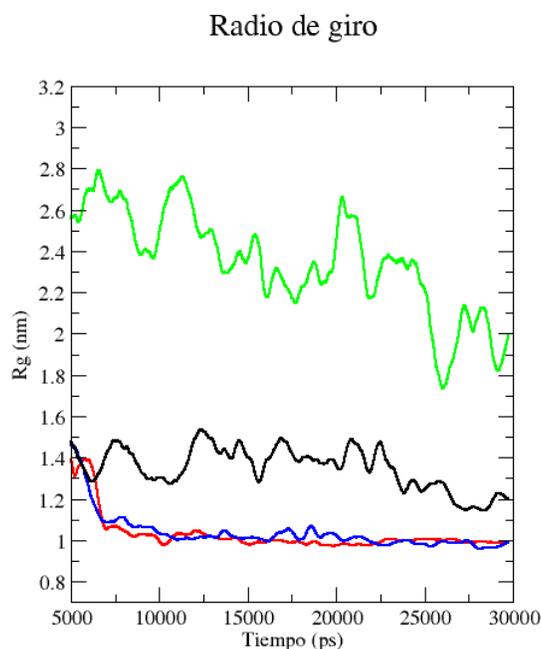


Figura 24. Radio de giro péptido 3

RMSF y Factor B

De la figura 25 se observa que la flexibilidad del péptido P3F aumenta en 30 de los aminoácidos con relación al P3. Las serinas, posiciones 50,51,53,54,58 y 59, presentan mayor movilidad al estar fosforiladas en el péptido fosforilado P3F, respecto su forma sin fosforilar en el péptido P3. La presencia de Ca^{2+} en el péptido P3Ca aumenta la movilidad de 29 aminoácidos en relación con el péptido P3. La movilidad de los aminoácidos del péptido P3FCa aumenta en comparación con el péptido P3, siendo mayor que el aumento mostrado por el péptido P3F. La Sep-50 del péptido P3FCa presenta una disminución de movilidad con respecto al péptido P3F.

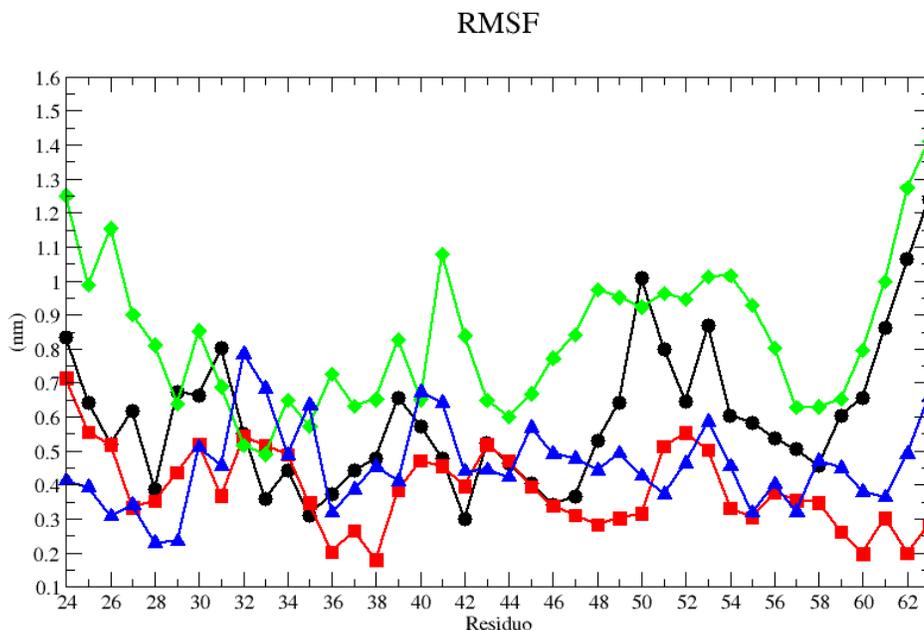


Figura 25. RMSF péptido 3

La figura 26 muestra las estructuras promedio a lo largo de las simulaciones de DM, así como los factores B, de las distintas versiones del péptido 3 (P3, P3F, P3Ca y P3FCa). Se observa que el péptido P3FCa (Figura 26 D) es la versión con la mayor flexibilidad, así como una estructura promedio completamente abierta. El péptido P3F (Figura 26 B) presenta un incremento de flexibilidad, observado por la presencia de colores verdes, amarillos y tonalidades claras de azul, con respecto al péptido P3 (Figura 26 A) donde predomina el color azul oscuro. También es notable una estructura menos compacta. La presencia de iones Ca^{2+} en el péptido P3Ca (Figura 26 C) tiene como consecuencia un ligero aumento de flexibilidad en algunas regiones de la cadena peptídica, en comparación al péptido P3, identificables por una tonalidad azul claro.

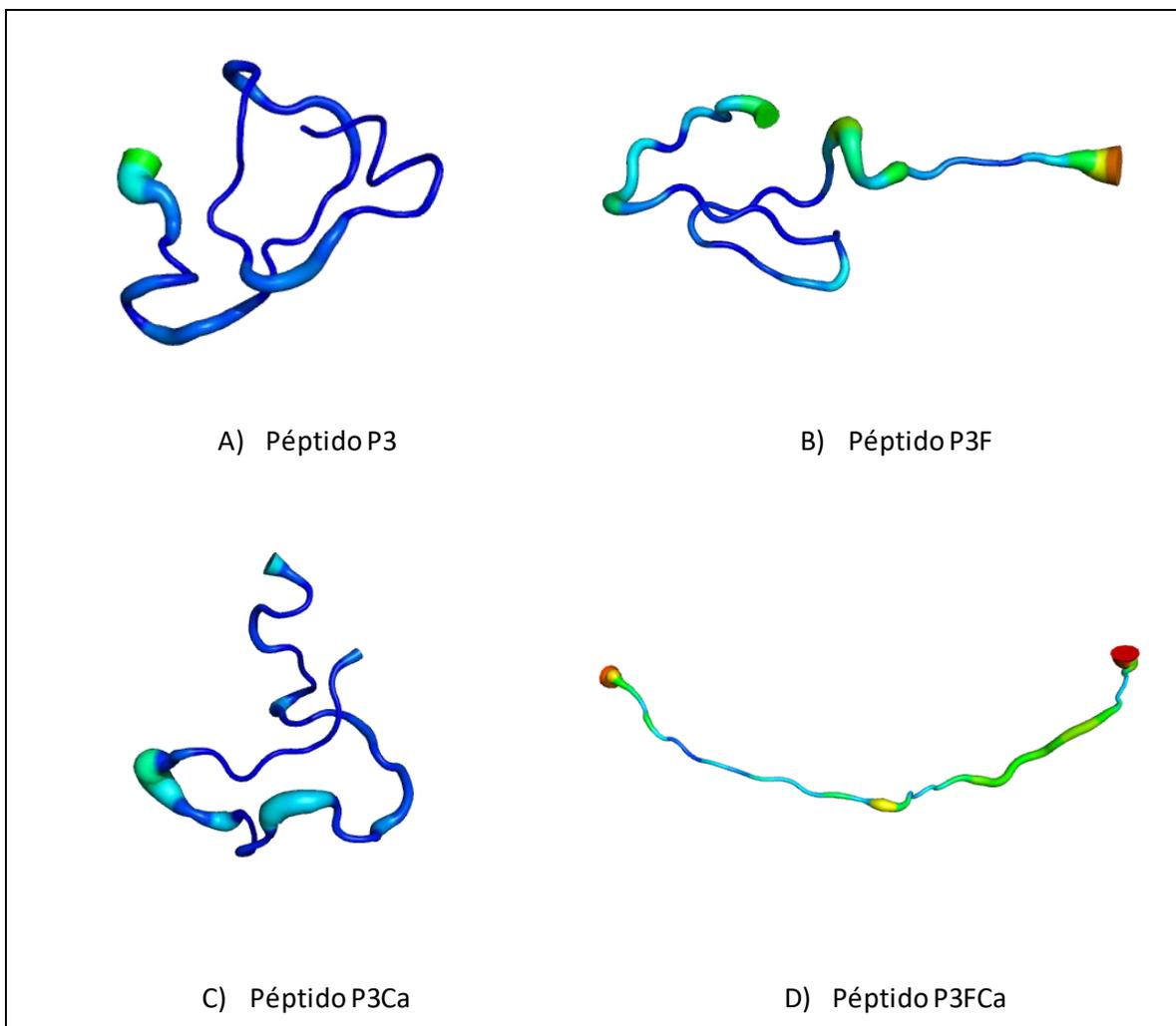


Figura 26 Estructura promedio de las diferentes versiones del péptido 3 y factor B representado por un espectro de colores azul-rojo

PH

De la tabla 10, se observa que la presencia de fosforilaciones en el péptido P3F tiene un fuerte efecto sobre la formación de puentes de hidrogeno. El péptido P3F presenta la formación de 35 PH, de los cuales 26 involucran al menos una serina. Sin embargo, sólo 3 puentes formados tienen una duración mayor al 50% del tiempo de simulación. En el péptido P3, siete de los diecinueve puentes formados involucran serinas, únicamente dos presentan una duración mayor al 50%. Mientras que para el péptido P3Ca, sólo cuatro puentes involucran a la serina, todos con una duración menor al 40% del tiempo de simulación. El péptido P3FCA muestra una

disminución considerable en el número de puentes formados, en donde sólo se registran 3 PH, dos de ellos formados exclusivamente por serinas.

Tabla 12. Puentes de Hidrogeno formados en las distintas versiones del péptido 3

Péptido	Puentes formados (Donador – Aceptor)	Aparición del puente durante la dinámica (%)
P3	Glu60 – Asp56	76.0
	His38 – Glu62	70.2
	Leu39 – Glu62	66.1
	Tyr36 – Ser50	65.3
	Glu61 – Asp56	63.6
	Glu62 – Ser58	55.3
	Gly46 – Glu63	46.3
	Ser51 – Tyr34	41.3
	Lys40 – Glu62	40.5
	Phe42 – Gln45	35.5
	Tyr34 – Tyr28	33.0
	Gln45 – Glu61	32.2
	Ser59 – Asp49	30.5
	Leu29 – Asp49	29.7
	Arg26 – Tyr58	28.9
	Ser59 – Asp46	28.1
	Tyr30 – Asp49	28.1
Ser59 – Asp56	27.2	
Ser59 – Asp56	25.6	
P3F	Phe42 – Gln45	82.6
	Sep48 – Sep47	70.2
	Lys40 – Pro37	68.6
	Sep47 – Sep47	64.4
	His38 – Glu52	64.4
	Val44 – Tyr28	63.6
	Gln45 – Sep47	55.3
	His38 – Sep48	48.7

P3F	His38 – Glu52	45.4
	Arg11 – Sep47	43.8
	Arg24 – Sep51	42.9
	His38 – Glu52	42.1
	His38 – Glu52	41.3
	Sep59 – Sep59	41.3
	Tyr28 – Tyr34	38.0
	Arg41 – Sep47	38.0
	Sep47 – Sep47	38.0
	Arg26 -Sep51	36.4
	Arg24 – Sep50	35.5
	Arg24 – Sep51	33.9
	Arg24 – Sep51	33.0
	Arg41 – Sep47	33.0
	Arg26 – Sep51	33.0
	Arg61 – Sep51	32.2
	Gln45 – Sep47	32.2
	Arg41 – Lys40	31.4
	Gly46 – Sep47	30.5
	Arg26 – Sep51	30.5
	Arg24 – Sep50	28.9
His38 – Sep48	28.1	
Arg26 – Sep51	27.2	
Sep51 – Sep51	27.2	
Arg41 – Sep47	26.4	
Arg24 – Sep51	26.4	
Arg26 – Sep51	26.4	
P3Ca	Phe35 – Lys31	77.6
	Tyr36 – Lys31	73.5
	Glu61 – Asp57	53.7
	Glu60 – Asp57	53.7
	Glu60 – Asp57	42.9
	Glu61 – Asp57	38.0

P3Ca	Arg26 – Ser58	38.0
	Glu62 – Asp57	36.3
	Glu62 – Asp57	36.3
	Tyr27 – Ser51	34.7
	Asn54 – Glu52	31.4
	Ser50 -Arg24	26.4
	Gly55 – Glu52	26.4
	Tyr36 – Ser46	26.4
	Glu53 – Glu52	25.6
	Lys31 – Tyr34	25.6
Tyr28 – Lys40	25.6	
P3FCa	Sep59 – Sep37	54.5
	Val44 – Lys18	52.0
	Sep51 – Sep29	41.3

Del análisis de calcios, se tiene lo siguiente:

Tabla 13. Distancia entre residuos y iones Ca²⁺ formadores de complejo

Péptido	Ca²⁺ en interacción	Residuo con el que interactúa	Distancia inicial COM residuo – COM Ca²⁺ (nm)	Distancia mínima COM residuo – COM Ca²⁺ (nm)
P3FCa	6100	Sep58	5.914	0.862
	6102	Sep48	6.547	2.588
	6111	Sep50	5.192	2.292
	6113	Sep59	4.714	2.015
	6115	Sep51	4.706	1.799
	6131	Sep48	5.812	1.991
	6131	Sep50	5.939	2.292
	6134	Sep58	3.788	0.552
	6134	Sep59	3.674	0.361
	6143	Sep58	5.669	1.854
	6149	Sep47	4.655	0.780
	6154	Sep47	3.995	1.025
	6154	Sep50	4.053	1.462

Estructura secundaria

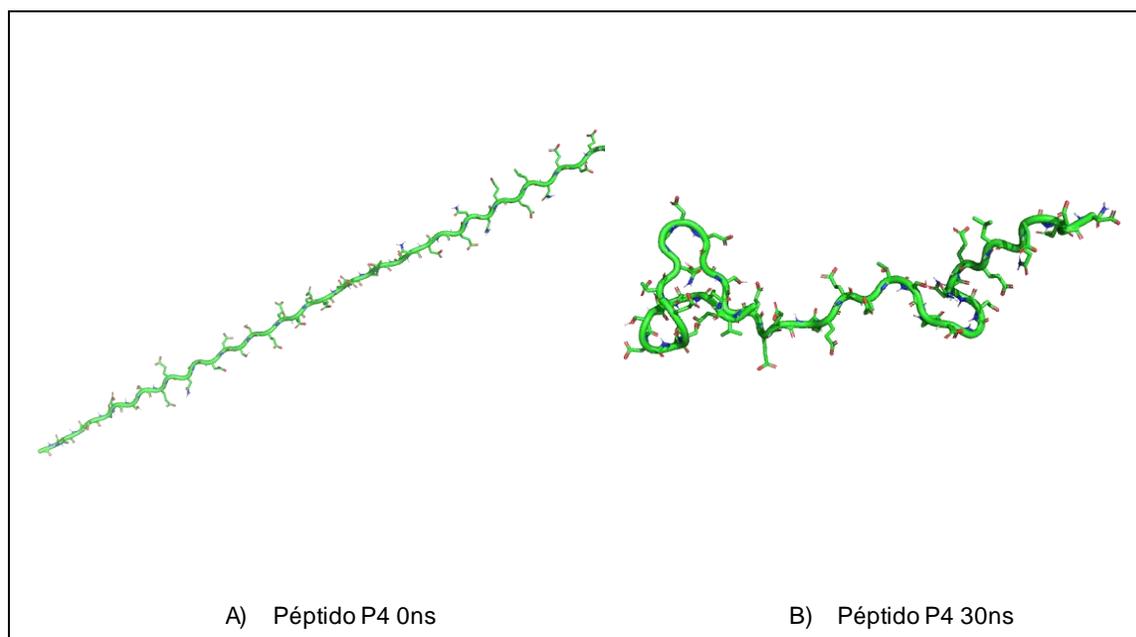
En la tabla 14 se muestran las estructuras secundarias determinadas para las diferentes versiones del péptido 3:

Tabla 14. Determinación de estructuras secundarias en el péptido 3

Péptido	DSSP		
	Estructura determinada		
	Hélice α	Hoja β	Giros
P3	13%	36%	51%
P3F	0%	35%	65%
P3Ca	0%	45%	55%
P3FCa	0%	22%	78%

5.2.4 Péptido 4

A continuación, las imágenes correspondientes a las estructuras iniciales y finales del péptido 4 en sus diferentes versiones (P4, P4F, P4Ca y P4FCa).



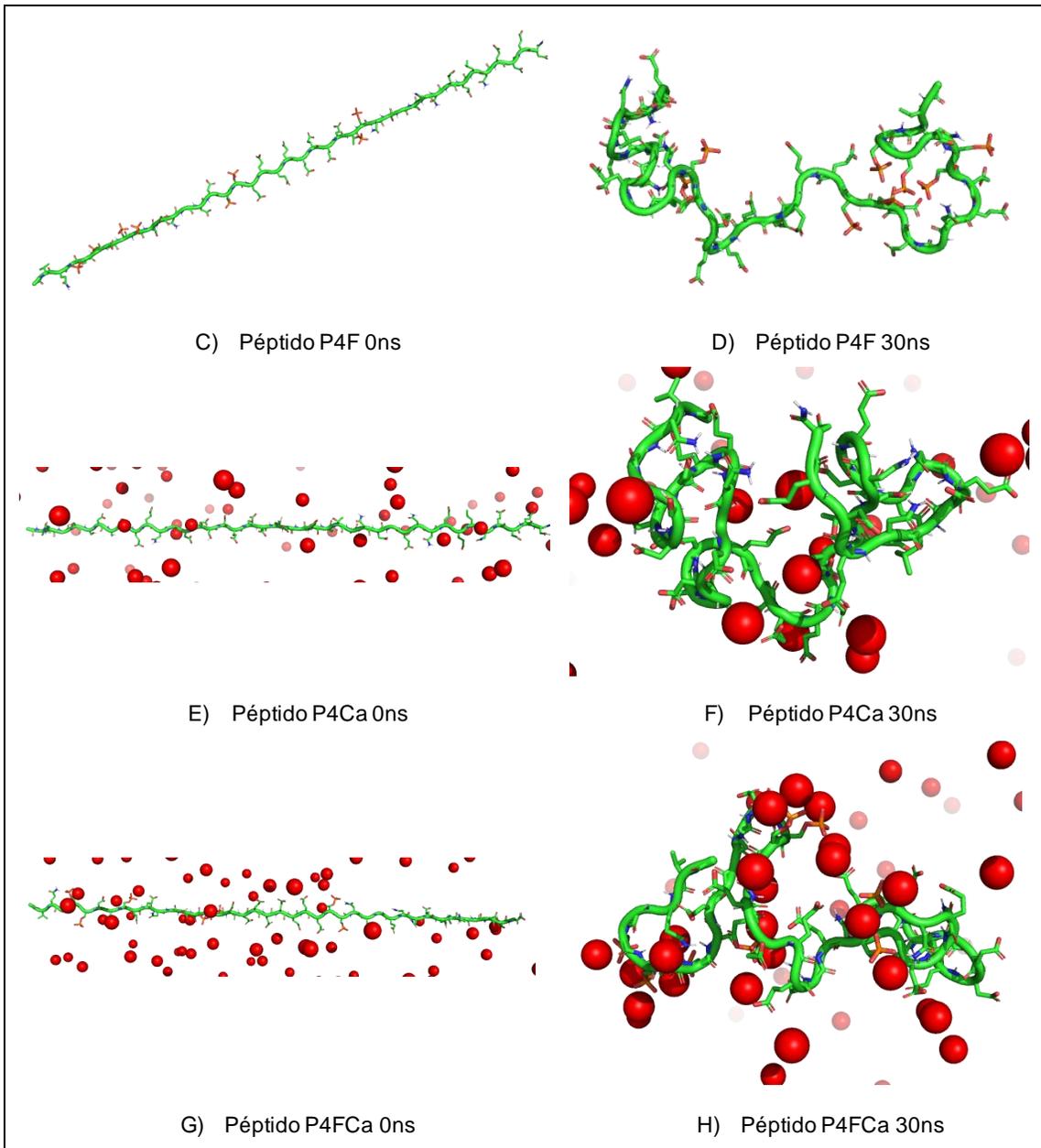


Figura 27 Estados estructurales de las distintas versiones del péptido 4 (Tabla 6)

La columna izquierda muestra el estado inicial, mientras que la columna derecha muestra las estructuras adquiridas al transcurrir los 30ns de simulación

Se observa que los péptidos P4 y P4F (Figura 25 B y D) presentan estructuras abiertas. Mientras que los péptidos P4Ca y P4FCa (Figura 25, F y H) tienden a formar estructuras con giros y bucles.

Rg

En la figura 28 se observa que no hay una tendencia clara en el comportamiento estructural de este péptido. No obstante, se aprecia que después de los 17.5ns la compactación entre los péptidos P4 y P4F es similar siendo, al final de la simulación, menor para el péptido P4. Los péptidos en presencia de iones Ca^{2+} , péptidos P4Ca y P4FCa, tienen una mayor compactación que el P4.

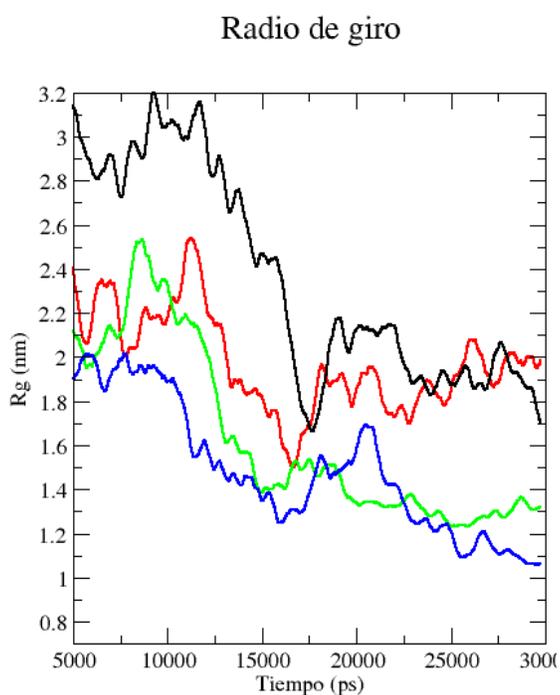


Figura 28 Radio de giro péptido 4

RMSF y Factor B

A continuación, se observan los resultados de flexibilidad de los péptidos:

RMSF

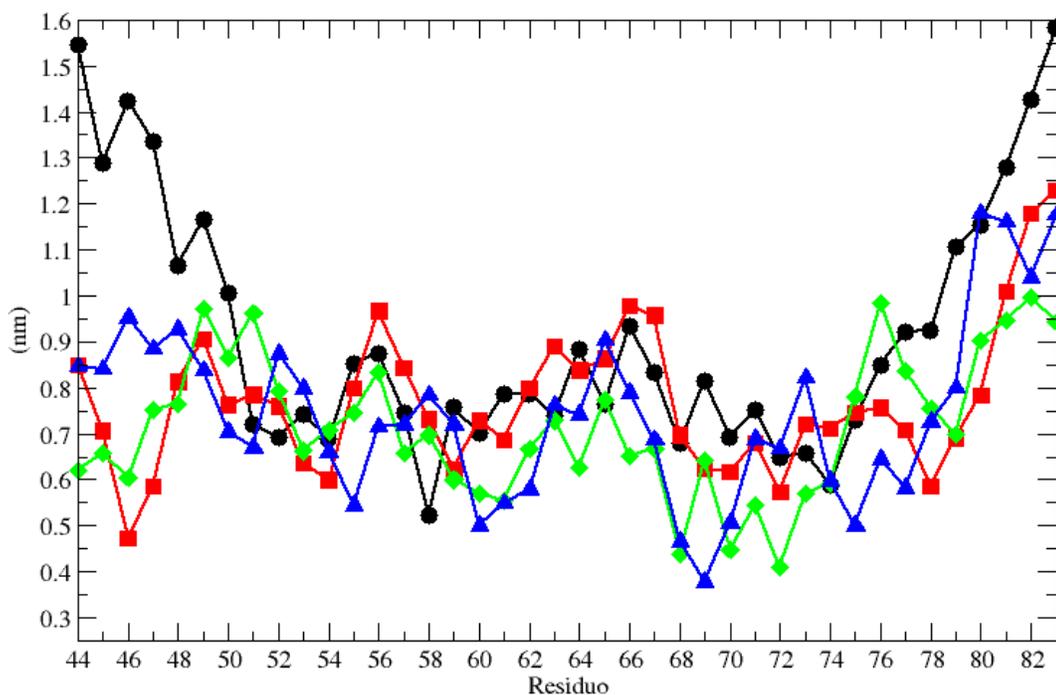
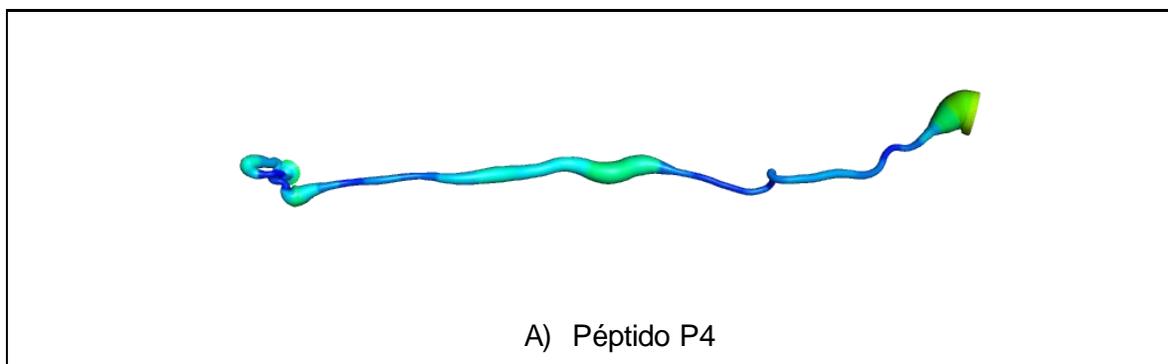


Figura 29 RMSF péptido 4.

De la figura 29 se observa un comportamiento similar para gran parte de la secuencia peptídica, para las cuatro versiones del péptido 4 (P4,P4F, P4Ca y P4FCA) donde la mayor diferencia se encuentra en los primeros siete aminoácidos, por parte del péptido P4F. A continuación, la figura 30 con los respectivos factores B:



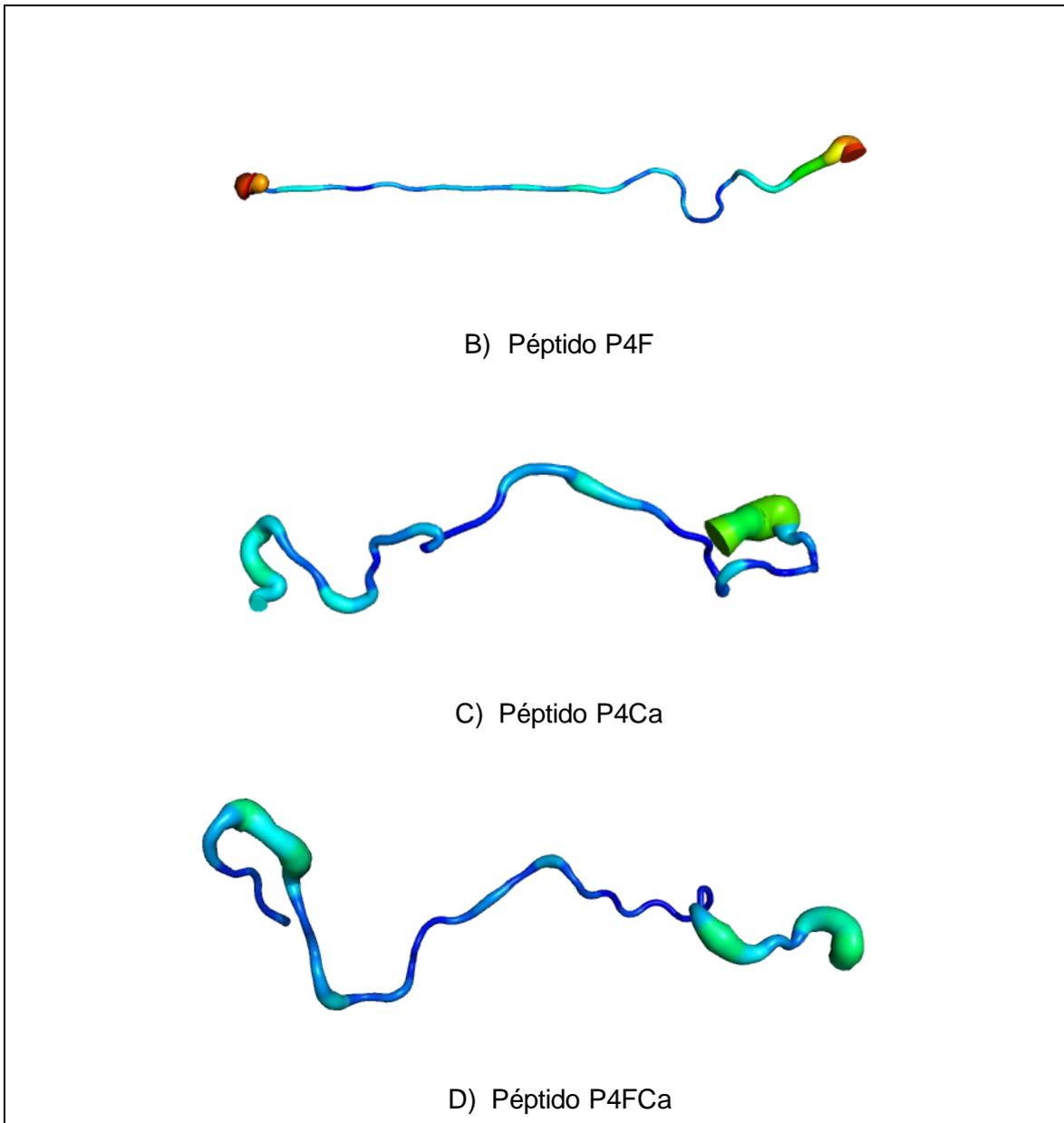


Figura 30. Estructura promedio de las diferentes versiones del péptido 4 y factor B representado por un espectro de colores azul-rojo.

Se observa que los péptidos tienden a formar estructuras abiertas, y que los cambios de flexibilidad se presentan principalmente en los extremos de los péptidos (Figura 30 A-D), correspondientes a las secuencias VQGSSDSS (extremo amino terminal) y EELNEDED (extremo carboxilo terminal). El péptido PF4 presenta la mayor flexibilidad para ambas regiones seguido por el péptido P4Ca,

posteriormente el péptido P4 y por último el péptido P4FCa con la menor flexibilidad de las cuatro versiones.

PH

El análisis de PH se resume en la tabla 15

Tabla 15. Puentes de Hidrogeno formados en las distintas versiones del péptido 4

Péptido	Puentes formados (Donador – Aceptor)	Aparición del puente durante la dinámica (%)
P4	Gly46 – Glu52	51.2
	Asn74 – Ser69	39.6
	Gln45 – Ser59	30.5
	Glu71 – Glu71	30.5
	Asn75 – Asn70	27.2
	Gln45 – Asp56	26.4
	Gly55 – Gln45	26.4
P4F	Sep47 – Sep47	67.7
	Sep48 – Sep47	66.1
	Sep69 – Tpo68	45.4
	Leu78 – Asn74	42.1
	Asn70 – Tpo68	35.5
	Asn70 – Tpo68	34.7
	Asn75 – Tpo68	31.4
	Glu77 – Asn74	30.5
	Sep60 – Tpo68	27.2
P4Ca	Glu61 – Asp57	63.6
	Ser59 – Asp57	36.3
	Ser59 – Asp57	36.3
	Ser50 – Ser47	35.5
	Ser59 – Asp57	28.9
	Gln45 – Ser50	28.1
P4FCa	Asn70 – Tpo68	53.7
	Val44 – Glu53	52.0

P4FCa	Asn70 – Tpo68	51.2
	Sep69 – Sep69	50.4
	Glu53 – Val44	42.1
	Sep69 – Tpo68	42.1
	Glu77 – Asn74	39.6
	Asn70 – Tpo68	39.6
	Tpo68 – Asn75	37.2
	Glu73 – Asn70	31.4
	Sep50 – Gln45	28.1
	Asn70 – Tpo68	27.2
Asp57 – Asn54	26.4	

Se observa que la treonina fosforilada tiene mayor propensión a formar puentes de hidrógeno que las serinas fosforiladas. A diferencia de lo que se ha observado hasta ahora, la treonina no disminuye tan abruptamente su capacidad de formar puentes de hidrógeno al interactuar con Ca^{2+} , a diferencia de la serina. Al ser la única treonina fosforilada presente en los péptidos, no será posible realizar más comparaciones al respecto.

Por último, para la asociación de calcio se tiene lo siguiente:

Tabla 16. Distancia entre residuos y iones Ca^{2+} formadores de complejo

Péptido	Ca^{2+} en interacción	Residuo con el que interactúa	Distancia inicial COM residuo – COM Ca^{2+} (nm)	Distancia mínima COM residuo – COM Ca^{2+} (nm)
P4FCa	7947	Sep47	3.339	2.408
	7947	Sep48	3.344	2.211
	7958	Sep59	4.338	0.689
	7974	Sep51	3.402	1.412
	7979	Sep58	3.797	3.990
	7980	Sep50	3.681	3.681
	8005	Sep51	4.559	0.625
	8014	Sep48	5.250	2.958

P4FCa	8014	Sep50	5.322	2.824
	8036	Sep42	4.064	2.641
	8037	Sep59	4.234	2.832
	8049	Sep69	2.960	1.512
	8069	Tpo68	4.622	3.690
	8076	Tpo68	2.525	2.316
	8076	Sep69	2.657	2.402
	8094	Sep47	5.202	1.538

Estructura secundaria

A continuación, se muestran las estructuras secundarias determinadas para las distintas versiones del péptido 4.

Tabla 17. Determinación de estructuras secundarias en el péptido 4

Péptido	DSSP		
	Estructura determinada		
	Hélice α	Hoja β	Giros
P4	1%	42%	57%
P4F	3%	32%	65%
P4Ca	2%	53%	45%
P4FCa	2%	43%	55%

5.2.5 Péptido 5

A continuación, en la figura 31 A-H, se muestran las imágenes correspondientes a los estados estructurales iniciales y finales de las distintas versiones del péptido 5:

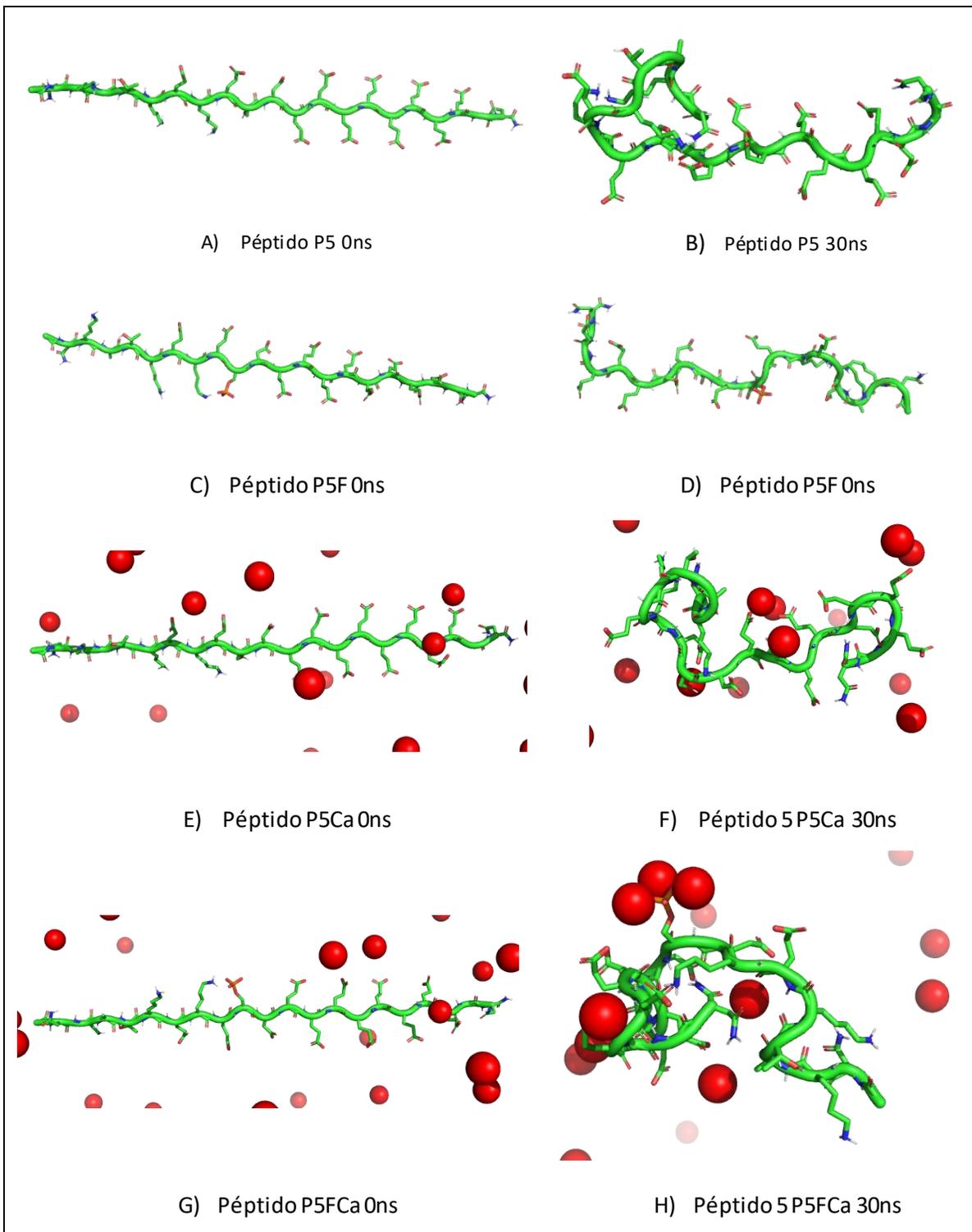


Figura 31. Estados estructurales de las distintas versiones del péptido 5 (Tabla 6). La columna izquierda muestra el estado inicial, mientras que la columna derecha muestra las estructuras adquiridas al transcurrir los 30ns de simulación.

De la Figura 31 B y D se observa que los péptidos P5 y P5F presentan una estructura similar a sus estados iniciales correspondientes (figura 31 A y C). Mientras que los péptidos P5Ca y P5FCa (figura 29 F y H) presentan formación de *loops* en su estructura.

Rg

La comparación de compactación estructural entre las diferentes versiones del péptido 5 (P5, P5F, P5Ca y P5FCa) se realiza con el gráfico de Rg mostrado en la figura 32.

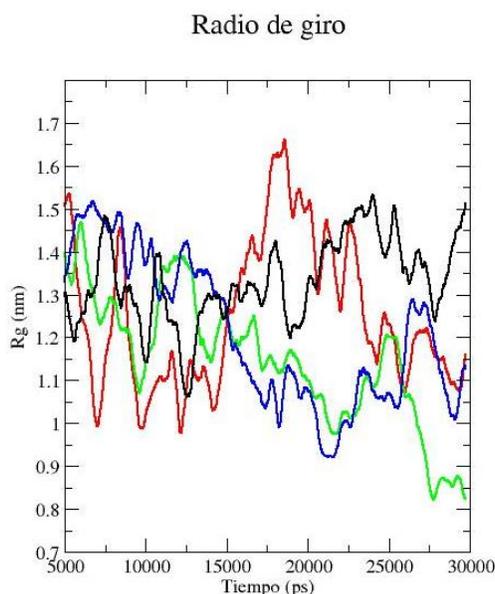


Figura 32. Radio de giro péptido 5.

Sin embargo, se observa que estos no siguen patrón alguno, sólo se puede decir que, para el punto final de la simulación los péptidos P5F y P5FCa representan ambos extremos del gráfico, con la mayor y la menor compactación, respectivamente. Mientras que los péptidos P5 y P5Ca muestran un grado de compactación similar. Los Rg promedio, de mayor a menor, son: P5F 13.30 Å, P5 12.52 Å, P5Ca 12.12 Å y P5FCa 11.49 Å.

RMSF y Factor B

De la figura 33 se observa que la fosforilación de la S-133 en el péptido P5F, provoca una disminución de movilidad en trece aminoácidos, respecto al péptido P5. Mientras que la presencia de iones Ca^{2+} , en conjunto a la fosforilación de la S-133, también disminuye la movilidad de quince aminoácidos, en comparación con el péptido P5, teniendo valores similares a los presentados por el P5F. La presencia de iones Ca^{2+} en el péptido P5Ca tiene como efecto la disminución de la movilidad en los aminoácidos al compararse con el P5.

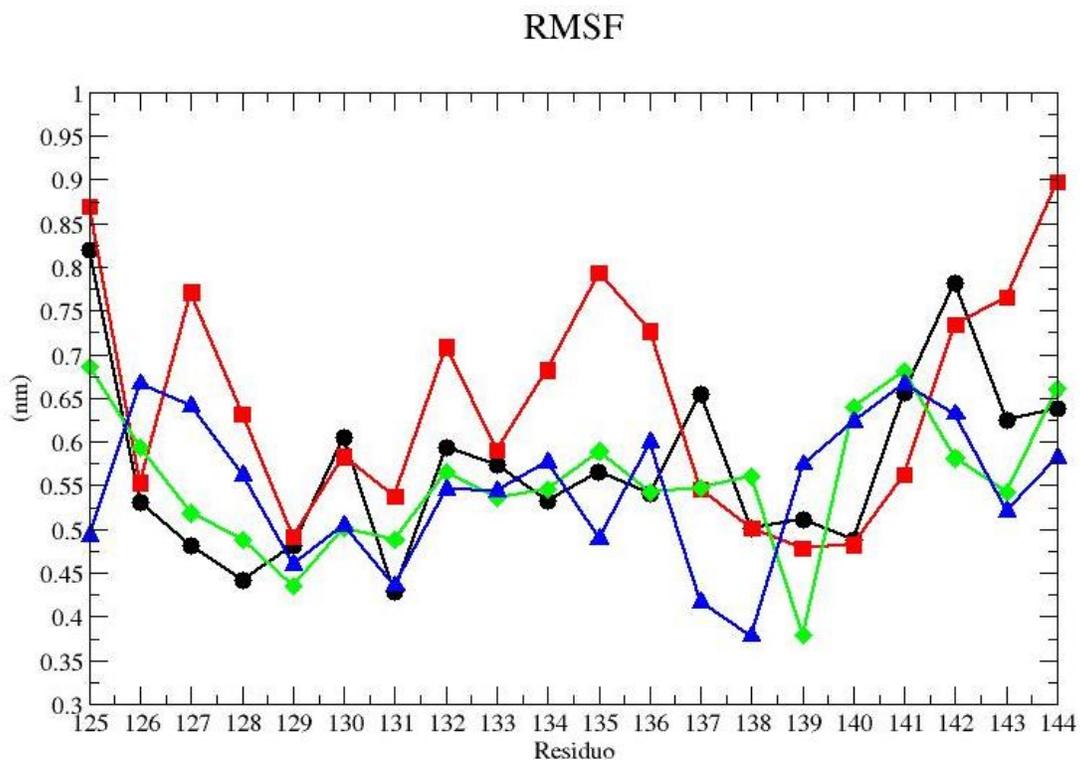


Figura 33 RMSF péptido 5

A continuación, las estructuras promedio junto al factor B de las versiones del péptido 5 (P5, P5F, P5Ca y P5FCa):

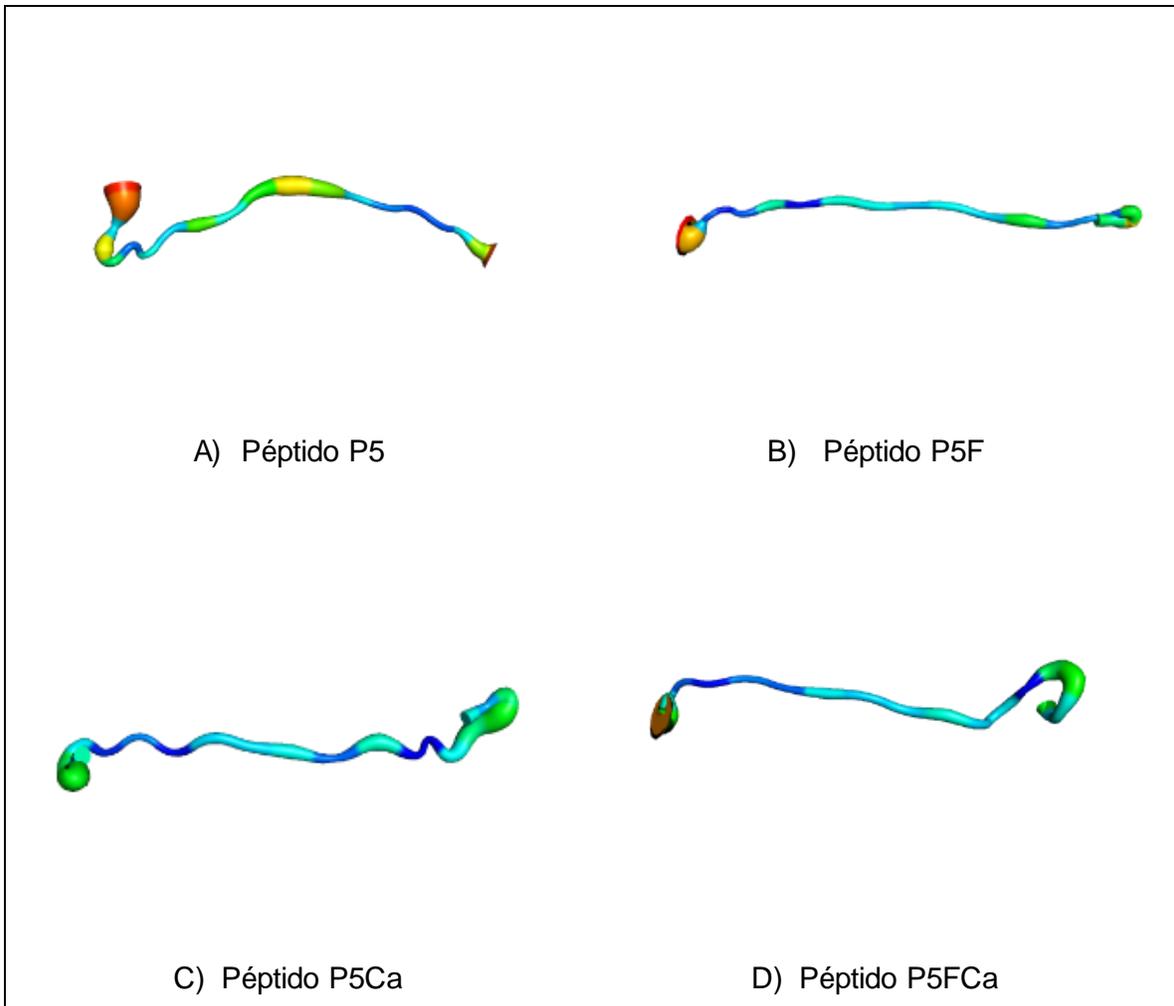


Figura 34. Estructura promedio de las diferentes versiones del péptido 5 y factor B representado por un espectro de colores azul-rojo

En la figura 34 se observa que las diferentes versiones del péptido 5 tienen una estructura promedio abierta y poco compacta. También es posible apreciar la diferencia de flexibilidades entre péptidos: el péptido P5 presenta la cadena más flexible al tener presencia de colores verdes, amarillos y naranjas a lo largo de su estructura, en contraste con los colores azules mostrados por el resto de los péptidos. Mientras que los péptidos P5F, P5Ca y P5FCa presentan flexibilidades similares.

PH

A continuación, se muestran los PH formados:

Tabla 18. Puentes de Hidrogeno formados en las distintas versiones del péptido 5

Péptido	Puentes formados (Donador – Aceptor)	Aparición del puente durante la dinámica (%)
P5	Lys131 – Lys126	42.1
	Glu130 – Lys126	35.5
	Asn125 – Lys129	30.6
P5Ca	Thr128 – Asn125	32.0
P5FCa	Glu137 – Asp134	31.4
	Gly143 – Glu139	28.9
	Glu142 – Glu139	27.2
	Glu141 – Glu139	25.6

En la tabla 18, se observa que el péptido P5F es el único de los péptidos analizados en el que no hay una formación de puentes de hidrogeno que perduren por más de un 25% del tiempo total de la simulación. A la vez, se observa que no hay participación de la Sp-133 en la formación de puentes de hidrógeno para el péptido P5FCa, pero hay glutamatos involucrados en todos los puentes detectados.

La tabla 19 muestra la información sobre las interacciones aminoácido – iones Ca^{2+} :

Tabla 19. Distancia entre residuos y iones Ca^{2+} formadores de complejo

Péptido	Ca^{2+} en interacción	Residuo con el que interactúa	Distancia inicial COM residuo – COM Ca^{2+} (nm)	Distancia mínima COM residuo – COM Ca^{2+} (nm)
Péptido 5 fosforilado	5493	Sep133	2.517	1.383
	5512	Sep133	3.757	0.981
	5527	Sep133	4.218	1.021

Estructura secundaria

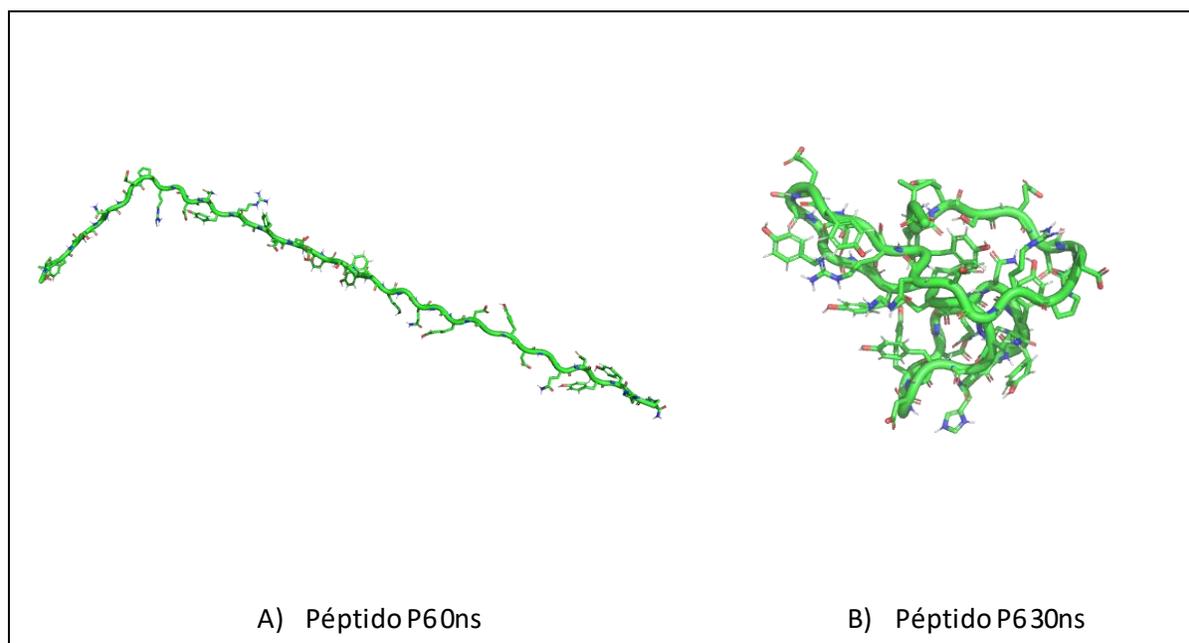
En la tabla 20 se muestran las estructuras secundarias determinadas para las distintas versiones del péptido 5.

Tabla 20. Determinación de estructuras secundarias en el péptido 5

Péptido	DSSP		
	Estructura determinada		
	Hélice α	Hoja β	Giros
P5	1%	26%	73%
P5F	0%	26%	74%
P5Ca	1%	39%	60%
P5FCa	3%	40%	57%

5.2.6 Péptido 6

A continuación, se presentan las imágenes respectivas para cada versión del péptido:



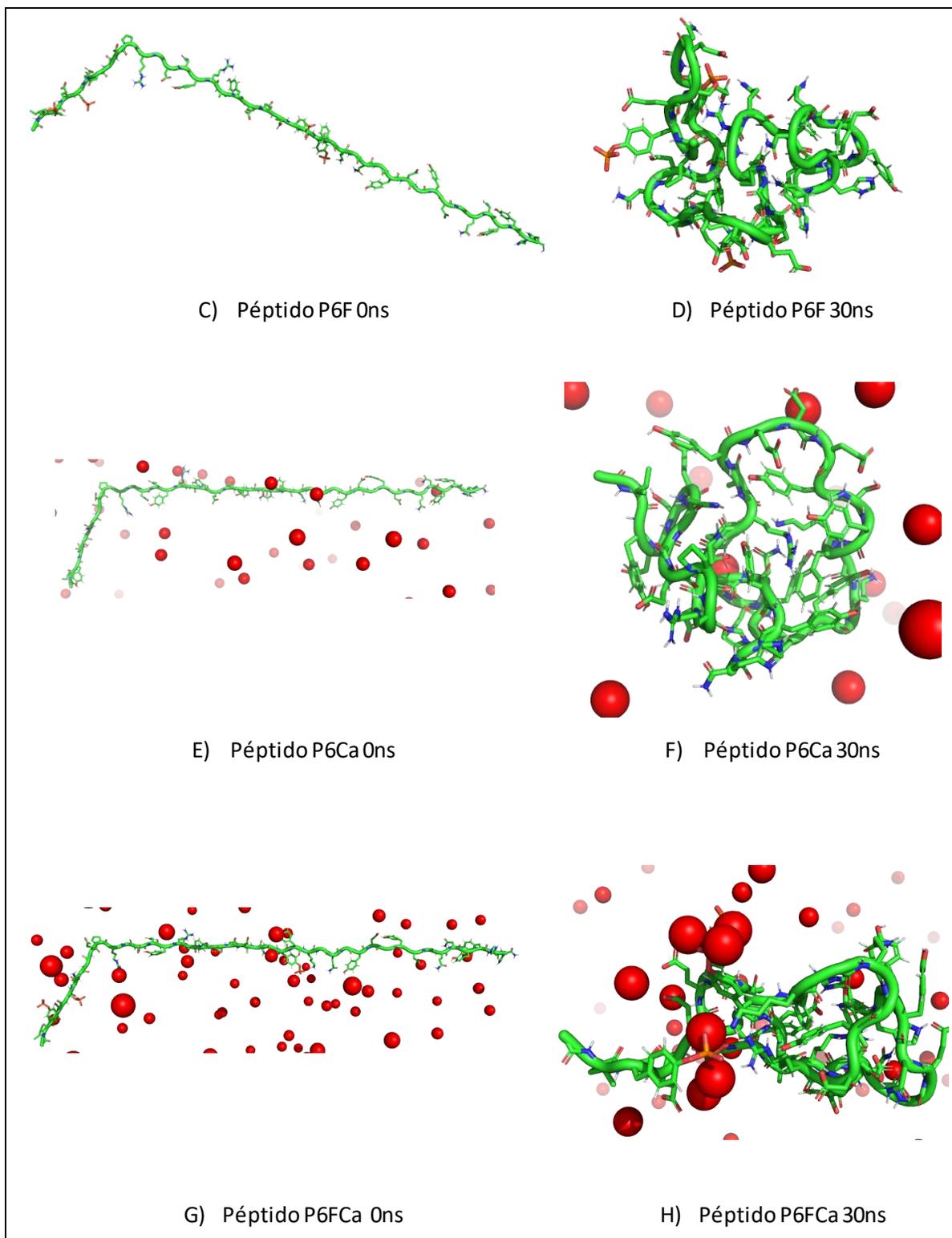


Figura 35. Estados estructurales de las distintas versiones del péptido 6 (Tabla 6)

La columna izquierda muestra el estado inicial, mientras que la columna derecha muestra las estructuras adquiridas al transcurrir los 30ns de simulación.

En la figura 33 (B,D,F y H) se observa que independientemente de la presencia de grupos fosfato o iones Ca^{2+} el péptido tiene una tendencia a formar giros y bucles.

Rg

Para poder estudiar mejor su conformación espacial, en la figura 36 se muestran los Rg de cada péptido. Se observa que los 4 péptidos muestran un comportamiento similar, con la tendencia a compactarse con el paso del tiempo, donde al final de la dinámica muestran valores cercanos de Rg que oscilan entre 9.5 y 10.5 Å.

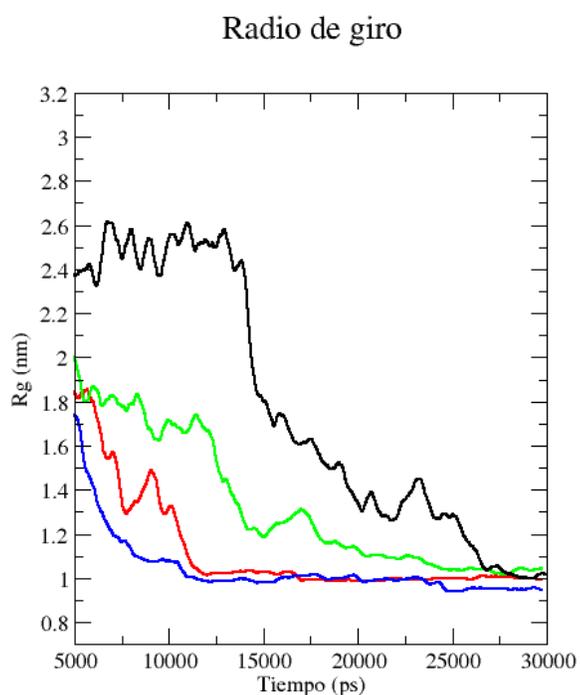


Figura 36. Radio de giro péptido 6.

RMSF y Factor B

A continuación, en la figura 37 se muestran los RMSF de cada péptido:

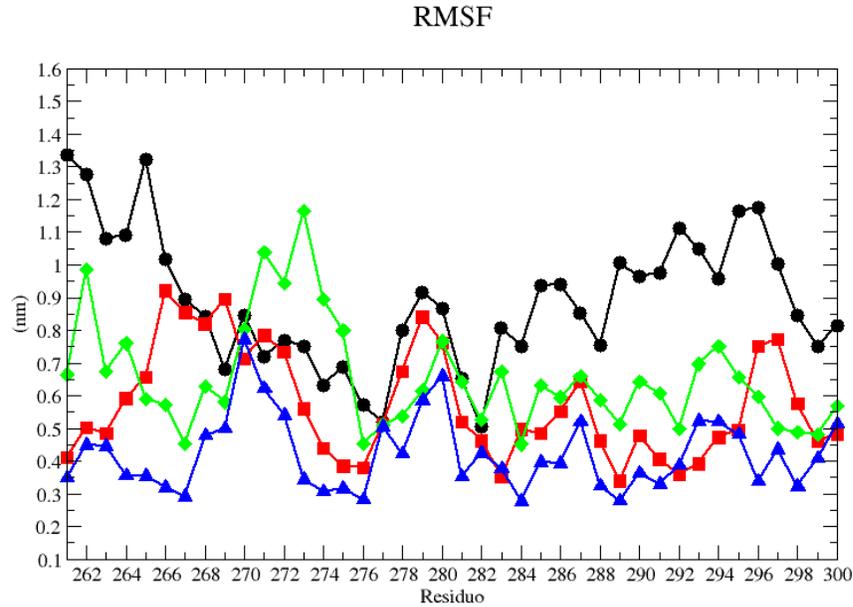
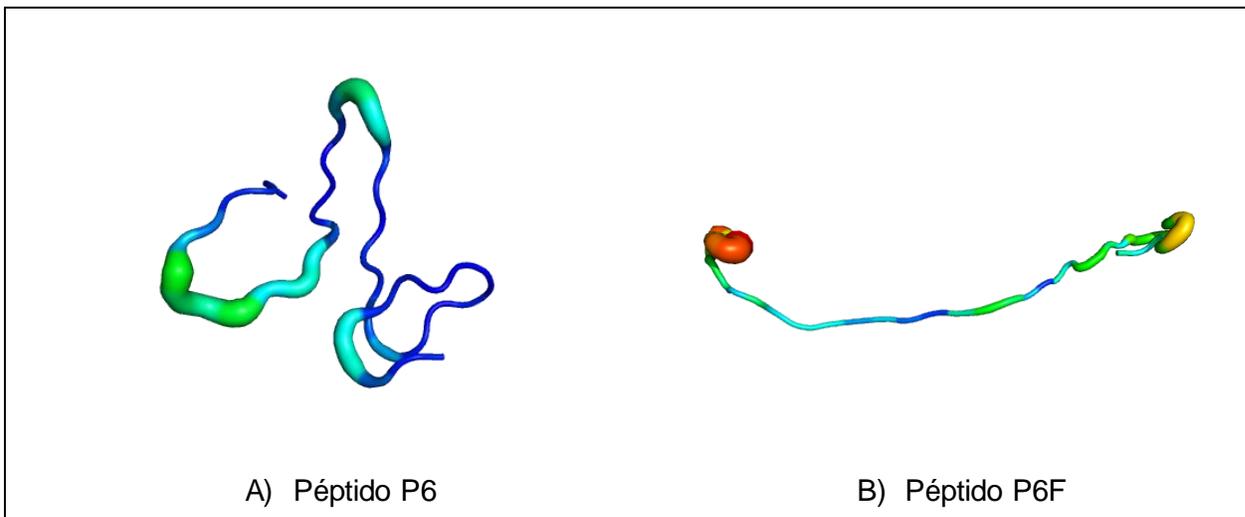


Figura 37. RMSF péptido 6.

En la figura 37 se observa que la tendencia general en la flexibilidad es P6F > P6FCa > P6 > P6Ca. La adición de grupos fosfato, péptido P6F, provoca un aumento en la movilidad de treinta y ocho aminoácidos de la secuencia del péptido P6. La adición de iones Ca^{2+} y de fosforilaciones en el péptido P6FCa tiene como efecto una disminución en la flexibilidad, con respecto a los valores mostrados por el péptido P6.



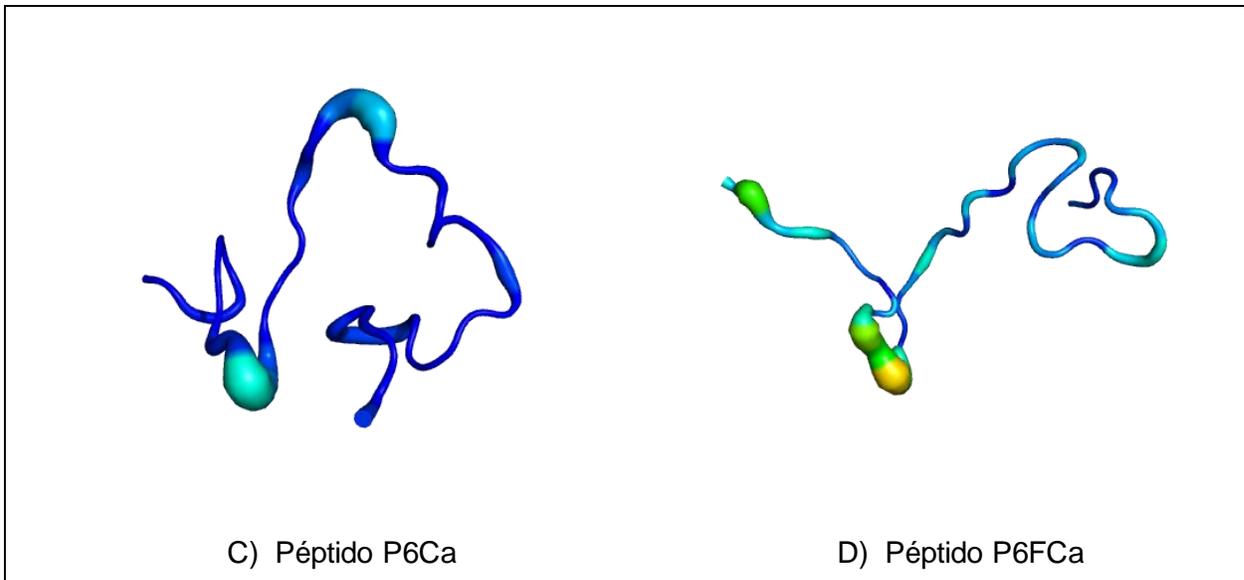


Figura 38. Estructura promedio de las diferentes versiones del péptido 6 y factor B representado por un espectro de colores azul-rojo

Se observa que el péptido P6F (Figura 38 B) tiene un aumento de flexibilidad global con relación al péptido P6 (Figura 38 A), lo cual se identifica por la presencia de colores amarillo, naranja, verde y tonalidades claras de azul, mientras que en el péptido P6 predomina una tonalidad azul oscuro. Del mismo modo, la presencia de iones Ca^{2+} y fosforilaciones en el péptido P6FCa (Figura 38 D) aumenta la flexibilidad en algunas regiones del péptido; sin embargo, este incremento es menor al mostrado por el P9F. La flexibilidad mostrada por el péptido P6Ca (Figura 38 C) es menor a la del péptido P6, se observa una predominancia de un color azul oscuro a lo largo de la estructura.

PH

En la tabla 21 se muestran los resultados del análisis de puentes de hidrogeno para el péptido 6.

Tabla 21. Puentes de Hidrogeno formados en las distintas versiones del péptido 6

Péptido	Puentes formados (Donador – Aceptor)	Aparición del puente durante la dinámica (%)
P6F	Glu268 – Sep264	58.1
	Sep264 – Sep264	56.7
	His300 – Asp290	49.6
	Arg270 – Sep264	48.9
	Arg270 – Sep264	48.9
	Glu265 – Sep264	46.0
	Gly267 – Sep264	43.9
	Phe284 – His300	43.2
	His300 – Asp290	41.1
	Tyr289 – Asp293	40.4
	Arg270 – Sep264	35.4
	Asn266 – Sep264	34.0
	Arg270 – Ile261	30.4
	Tyr297 – Gln287	28.3
	Asn296 – Gln287	28.3
Lys285 – Ptr283	26.2	
P6FCa	Glu278 – Gly267	48.7
	Tyr292 – Lys285	46.3
	Tyr397 – Gln295	43.8
	His299 – Tyr292	42.1
	Arg275 – Asp272	38.0
	Arg270 – Sep264	32.2
	Ptr283 – Glu265	31.4
	Arg270 – Sep264	28.9
	Sep264 – Sep264	28.0
	Arg270 – Sep264	27.2
	Arg270 – Sep264	27.2
	Gly267 – Ptr281	26.4
	Glu268 – Ser282	25.6

P6FCa	Arg275 – Asp272	25.6
P6	Tyr38 – Gly286	61.1
	Ala17 – Ser282	58.6
	Phe25 – Tyr274	57.0
	His40 – Asp293	54.5
	Gly29 – Tyr297	54.5
	Lys26 – Gly294	54.5
	Tyr30 – Gln295	51.2
	His299 – Asp293	50.4
	Gln295 – Asp293	33.8
	His300 – Gln295	33.8
	Gly267 – Tyr297	33.8
	Gln295 – Asp293	33.0
	Asn296 – Glu265	27.2
	Asn266 – Glu268	27.2
	Ser282 – Ala276	26.4
Asn296 – Phe284	25.6	
Glu265 – Glu268	25.6	
P6Ca	Glu267 – Asp272	67.7
	Ser264 – Tyr262	62.8
	His299 – Asp272	50.4
	Gly286 – Tyr292	47.1
	His300 – Arg270	47.1
	His299 – Asp290	46.2
	His299 – Asp290	43.8
	Glu263 – Gln295	42.1
	Asn266 – Ile261	40.4
	Phe284 – Tyr281	38.0
	Arg275 – Asn296	36.3
	His299 – Asp290	36.3
	His299 – Asp290	34.7
	Tyr274 – Asn266	34.7
Tyr292 – Tyr289	33.8	

P6Ca	Tyr281 – Lys285	33.0
	Tyr298 – Tyr283	30.5
	Arg270 – Glu268	28.9
	Glu265 – Ile261	28.0
	Lys285 – Ser282	27.2

A continuación, los resultados correspondientes al análisis de distancias entre iones Ca^{2+} y los péptidos.

Tabla 22. Distancia entre residuos y iones Ca^{2+} formadores de complejo

Péptido	Ca^{2+} en interacción	Residuo con el que interactúa	Distancia inicial COM residuo – COM Ca^{2+} (nm)	Distancia mínima COM residuo – COM Ca^{2+} (nm)
Péptido 9 fosforilado	4084	Ptr3	8.489	2.840
	4828	Sep5	8.993	2.942
		Ptr3	10.391	3.974
	4849	Sep5	10.615	3.422
		Ptr22	9.255	2.805
	4873	Ptr24	8.718	2.326
		Ptr22	10.516	2.504
	4913	Ptr24	8.718	2.326
		Ptr24	9.173	2.090

Estructura secundaria

En la tabla 23 se muestran las estructuras secundarias determinadas para las distintas versiones del péptido 6:

Tabla 23 Determinación de estructuras secundarias en el péptido 5

Péptido	DSSP		
	Estructura determinada		
	Hélice α	Hoja β	Giros
P6	0%	46%	54%
P6F	0%	44%	56%
P6Ca	4%	52%	44%
P6FCa	0%	43%	57%

6.0 Discusión

6.1 Selección de péptidos

Este trabajo forma parte de un proyecto cuyo propósito es identificar regiones de la BSP que pudieran estar involucradas en el proceso de nucleación de la HA. Sin embargo, debido a la gran cantidad de estudios que esto requiere, esta tesis se limita al estudio de las interacciones entre péptidos, con o sin fosforilaciones, y Ca^{2+} .

Como primer punto, fue necesario elegir distintos fragmentos de la BSP y analizar aquellos dominios de la secuencia que, de acuerdo con sus características estructurales, fisicoquímicas, conservación o modificaciones postraduccionales, puedan estar potencialmente involucrados en el proceso de biomineralización. Las consideraciones para discernir entre que fragmentos pueden ser de utilidad y cuáles no, están basadas en la información existente sobre la BSP y características funcionales de las IDP. Los fragmentos elegidos se consideran como péptidos individuales.

Las IDP, como la BSP, se caracterizan por no tener una estructura globular para interactuar con sus ligandos fisiológicos. Sin embargo, al interactuar con estos, puede existir adquisición de una estructura secundaria. También se sabe que las fosforilaciones pueden fomentar cambios estructurales en las IDP, así como favorecer algunas interacciones. Por ello, consideramos importante estudiar regiones desordenadas, posibles estructuras secundarias y zonas con presencia de fosforilaciones. La conservación de una secuencia, en concreto, puede significar que está implicada en el papel funcional de la proteína (Sitbon, 2007), por lo tanto, también se consideran aquellas regiones que presenten conservación entre especies.

Se sabe gracias a diversos estudios que las interacciones entre la BSP, así como otros miembros de la familia SIBLING, Ca^{2+} , colágeno y HA son de naturaleza electrostática (Tye, 2005, Vincent, 2013, Villarreal-Ramirez, 2014). Por lo cual, las zonas con mayor presencia de aminoácidos cargados positiva o negativamente y las fosforilaciones son zonas de gran interés en la selección. Los aminoácidos con

posiciones 47-83 (SSDSSEENGEENGDDSSSEEEEEEEETSNEGENNEEL NEDED) y 125-144 (NKATKEKESDEEEEEEEEGN) representan las zonas con mayor presencia de residuos cargados, donde cada una de éstas cuenta con la existencia de un octámero de E.

Otra zona de particular interés para esta discusión es la comprendida por los residuos 1-46. Esta región presenta una carga neta casi neutra en comparación con otras partes de la proteína; pero los aminoácidos 19-46 tienen uno de los mayores puntajes de conservación para el péptido (Figura 11), por lo tanto, se consideró para los estudios de dinámica molecular. El resto de los aminoácidos se encuentra flanqueando a una serina que ha sido reportada experimentalmente como fosforilada (sección 3.2.2) y se incluye para su análisis.

A continuación, se muestra una descripción de los péptidos seleccionados para su posterior estudio.

- Péptido 1 (4-23): Los resultados anteriores muestran que este péptido presenta una predicción para ser fosforilado en la S-15 (Tabla 2), siendo consistente con las determinaciones experimentales de aminoácidos fosforilados presentados por Zaia (2001). También se encuentra en una de las zonas más conservadas de la proteína. Además, muestra una predicción al desorden para las posiciones 4-17 (Tabla 3), correspondiente a un 70% del péptido. La predicción para la estructura secundaria según indican un péptido con giros aleatorios y la formación de hélice α para los residuos 5-9 y 4-20 (Tabla 4).
- Péptido 2 (24-43): Esta región de la proteína no presenta ninguna carga negativa, por el contrario, tiene aminoácidos cargados positivamente y aminoácidos hidrofóbicos; sin embargo, se ha demostrado que este dominio es importante para la interacción BSP – colágeno (Tye, 2005), esto aunado a su puntaje de conservación en el MSA es motivo por el cual se decidió utilizarlo para los estudios de simulaciones de DM.

- Péptido 3 (24-63): Este péptido comprende 15 residuos de aminoácidos correspondientes a glutamato, aspartato y serinas seleccionadas para ser fosforiladas (Tabla 2). Tiene una baja puntuación en conservación de la secuencia (Figuras 11-13), pero es una zona con alta presencia de aminoácidos capaces de tener interacciones electrostáticas. Presenta una región desordenada en los residuos 46-63 (Tabla 3). También cuenta con predicción para la formación de hélice α en su región terminal, correspondiente a glutamatos (Tabla 4).
- Péptido 4 (44-83): Este péptido contiene los últimos 20 aminoácidos del péptido 3, y el resto de los residuos conforman una de las zonas más ácidas de la proteína, lo que resulta de interés para analizar la unión de Ca^{2+} . Tiene predicción para IDR en su totalidad (Tabla 3), así como predicción para la formación de hélice en los residuos 60-66 correspondiente a las repeticiones de glutamato y 73-83, otra zona rica en glutamato (Tabla 4).
- Péptido 5 (125-144): Este fragmento contiene una secuencia consenso para fosforilación (S-133) por la CKII. Esta secuencia consenso se encuentra conservada en 25 de las 26 especies analizadas mediante MSA (Figura 12). Además, se ha demostrado su importancia biológica para el proceso de nucleación de cristales de HA en rata (Baht, 2010). La secuencia del P5 presento una predicción al desorden en un 100% (Tabla 3). Sin embargo, los programas de predicción de estructura secundaria determinaron la formación de hélice α en las repeticiones de glutamato (Tabla 4).
- Péptido 6 (261-300): Este péptido corresponde a la secuencia que rodea al dominio RGD. El MSA lo indica como una de las zonas con mayor conservación en secuencia (Figura 13). La predicción de estructura secundaria indica la formación de hélice α y giros aleatorios (Tabla 4). Los predictores indican desorden en algunos de sus aminoácidos (Tabla 3). Hay predicciones para fosforilaciones sobre Y-262, Y-281, Y-283 y S-264 (Tabla 2).

6.2 Simulaciones de DM

Se estudiaron distintas características de los péptidos a partir de los resultados obtenidos de las simulaciones de DM, con el propósito de identificar regiones que pudieran estar involucradas en el proceso de nucleación de HA.

El grado de compactación que adquieren los péptidos durante las simulaciones de DM se estudió mediante el análisis del Rg en distintas condiciones como la presencia de fosforilaciones y/o la adición de iones Ca^{2+} :

En los péptidos P1F, P3F y P5F se observa que la presencia de fosforilaciones disminuye el grado de compactación y aumentan su Rg. Mientras que en los péptidos P4F y P6F las fosforilaciones no parecen tener un efecto en el Rg. Chin (2016) demuestra que la presencia de fosforilaciones en una IDP puede inducir la extensión estructural sobre los alrededores de la modificación postraduccional. Esto es debido a efectos electrostáticos y estéricos generados por el grupo fosfato. Las interacciones electrostáticas rigen en gran medida el comportamiento de la BSP, por lo cual es muy probable que las interacciones entre grupos fosfato y aminoácidos con carga estén incrementando el Rg en los péptidos P1F, P3F y P5F, en comparación a los péptidos sin fosforilar P1, P3 y P5. Interesantemente, P1FCa y P5FCa tienen un Rg menor en comparación a P1F y P5F, respectivamente. La adición de iones Ca^{2+} resulta en una compactación estructural, probablemente por la disminución de la carga neta de los péptidos. Es importante destacar las tirosinas Y-281 y Y-283 fosforiladas del péptido 6, pues son las únicas fosforilaciones en zonas ordenadas de la proteína, de acuerdo con PONDR, PrDOS y Netphos.

Actualmente no hay un entendimiento claro sobre los detalles implicados en el mecanismo de nucleación de HA a pesar de que se han realizado múltiples estudios (Baht, 2008; Yang, 2010); sin embargo, el comportamiento exhibido por estos péptidos asemeja al mecanismo de interacción conocido como “*fly cast*”. El cual consiste en la apertura de una cadena peptídica para aumentar su radio de captura para promover enlaces a sitios específicos de sus ligandos (Shoemaker, 2000). Dicho mecanismo sería capaz de propiciar interacciones directas con iones Ca^{2+}

seguido de un rearrreglo estructural, pues se sabe que la unión de una IDP con su ligando provoca una transición estructural del tipo desorden-orden que puede manifestarse como una compactación del péptido (Bah, 2015). No obstante, el péptido P3FCa muestra una compactación menor a la del péptido P3F. Para este caso se consideran dos posibilidades, una de ellas es que se necesite más de un ligando para conferir un cambio estructural, lo que es consistente con lo propuesto por Fisher (2001) quien mencionan que la falta de orden en la estructura permite a la BSP interactuar de manera simultánea con diferentes moléculas, donde las interacciones con colágeno o los cristales recién nucleados de HA podrían inducir un cambio en este comportamiento. La segunda posibilidad es que a pesar de tener una interacción con un blanco biológico su estructura permanezca desordenada, fenómeno ya descrito en IDP por Kulkarni et al. (2017).

En los péptidos P1Ca, P3Ca, P5Ca el grado de compactación es similar al observado para los péptidos P1, P3 y P5, respectivamente. Estos péptidos no tienen fosforilaciones, pero sí tienen cargas negativas, lo cual implicaría que las fosforilaciones son necesarias para una modificación estructural. P2Ca muestra una compactación estructural mayor comparado con el péptido P2. Es probable que la presencia de aminoácidos cargados positivamente en interacción con los iones Ca^{2+} induzca un arreglo que permita una mayor compactación, influenciado por la presencia de aminoácidos hidrofóbicos.

El péptido 4, en sus distintas versiones, exhibe un comportamiento distinto al descrito para los péptidos anteriores. Las versiones con los iones Ca^{2+} (P4Ca y P4FCa) tienen un grado de compactación mayor al mostrado por el péptido P4, para explicar este fenómeno fue necesario un análisis con mayor detenimiento de las trayectorias correspondientes. Esto puede ser consecuencia de las interacciones electrostáticas entre los aminoácidos cargados negativamente, recordando que este péptido tiene una carga de -21 en su versión no fosforilada (P4) y de -37 en su versión fosforilada (P4F). La presencia de Ca^{2+} puede influir sobre la orientación de las cadenas laterales, permitiendo un acomodo más compacto del esqueleto

peptídico. Sin embargo, el péptido P4F muestra un comportamiento similar a P4, el cual adquiere una estructura extendida como consecuencia de su carga neta

El péptido 6 en todas sus versiones tiene un R_g similar, es decir, no hay variaciones mayores a 1 Å. Este péptido comprende la región del motivo RGD con funciones de adhesión celular e influencia sobre la concentración intracelular de Ca^{2+} , y se decidió estudiar por su importancia biológica y su conservación en secuencia. Sin embargo, lo observado gracias a las simulaciones de DM sugiere que su comportamiento es independiente de la presencia de calcio y/o de fosforilaciones; por lo tanto, este péptido podría no tener participación alguna en las interacciones BSP-HA-colágeno- Ca^{2+} .

Con respecto a los resultados del RMSF y de factores B, la movilidad de los péptidos se ve modificada debido a la presencia de las fosforilaciones y de los iones Ca^{2+} . Sin embargo, es difícil realizar un análisis residuo por residuo en presencia de tantas variables. Se sabe que la flexibilidad de una cadena está dada, en gran parte, por su secuencia de aminoácidos (Radivojac, 2004) por lo que cualquier interacción o modificación sobre la misma influirá en ella. Los cambios observados en la movilidad de los péptidos, asociados a su grado de compactación, permiten dividir a los péptidos estudiados en 6 grupos, todos ellos con respecto a las formas sin fosforilar y sin iones Ca^{2+} (P1, P2, P3, P4, P5 y P6):

- Aumento de movilidad y aumento del R_g : P1F, P3F y P3FCa.

Para este caso, se observa que los péptidos se encuentran menos compactos, en comparación con los péptidos P1 y P3, respectivamente. Esta mayor elongación favorece el libre movimiento de la cadena peptídica, al no haber impedimentos estéricos.

- Aumento de movilidad y R_g similar: P1Ca, P3Ca, P4F, P6F y P6FCa.

La compactación de los péptidos es similar a la de sus versiones sin fosforilar y sin iones Ca^{2+} ; sin embargo, presentan un aumento de movilidad en distintas regiones de la cadena peptídica, estos cambios son atribuidos a la modificación del péptido

original, ya sea con la adición de fosforilaciones, la adición de Ca^{2+} o ambas. Es probable que el aumento de movilidad observado en los péptidos P1Ca, P3Ca y P6FCa tenga como origen las interacciones electrostáticas entre iones Ca^{2+} y aminoácidos cargados positiva o negativamente. Para el péptido P4F la presencia de fosforilaciones en una estructura con alta carga negativa puede propiciar un mayor efecto de repulsión, provocando un reacomodo que permita disminuir estas interacciones al máximo para adquirir mayor estabilidad (Kohn, 2008). Es probable que un fenómeno similar ocurra en el péptido P6F, donde los aminoácidos fosforilados se encuentran rodeados de aminoácidos polares y cargados negativamente.

- Aumento de movilidad y disminución del Rg: P2Ca y P4Ca.

Estos péptidos presentan una mayor compactación que los péptidos P2 y P4, respectivamente, sin embargo, tienen mayor movilidad. El péptido P2 es el único en estudio con una carga neta positiva, por lo cual puede dar lugar a repulsiones electrostáticas en presencia de iones Ca^{2+} teniendo consecuencias sobre la movilidad del péptido. Para el péptido P4Ca los mayores cambios ocurren en los extremos de la cadena, probablemente debido a un acomodo que permita la interacción entre aminoácidos fosforilados y iones Ca^{2+} .

- Disminución de movilidad y aumento del Rg: P1FCa y P5F.

La disminución de movilidad en el P1FCa puede tener origen en las interacciones entre la serina fosforilada Sp-15 y iones Ca^{2+} , las cuales condicionan la movilidad al formar un complejo con este aminoácido, teniendo como consecuencia un descenso de ésta en los alrededores de la S-15. El péptido P5F presenta una región cargada negativamente que corresponde a más del 50% de la secuencia este péptido, es posible la adquisición temprana de una estructura elongada de mayor estabilidad, explicando así el decremento en la movilidad.

- Disminución de movilidad y Rg similar: P5Ca y P6Ca.

El péptido P5Ca, a pesar de tener un grado de compactación similar al péptido P5, muestra una disminución en su movilidad. Este péptido cuenta con un dominio con alta afinidad por iones Ca^{2+} , estos iones interactúan alrededor de dicho dominio, y posiblemente ocasionan baja movilidad al estabilizar las cadenas laterales de los glutamatos presentes en el péptido. Mientras que la disminución de la movilidad del péptido P6Ca puede deberse a una temprana estabilización temprana inducida por la presencia de iones Ca^{2+} .

- Disminución de movilidad y disminución del Rg: P4FCa y P5FCa.

Se observa que ambos péptidos presentan una mayor compactación que los péptidos P4 y P5, respectivamente. La explicación más plausible es la interacción entre péptidos con alta carga negativa y los iones Ca^{2+} resultando en una estructura más compacta comparada a su versión sin fosforilar y sin calcio.

El estudio de las estructuras secundarias formadas por los distintos péptidos tiene como propósito detectar aminoácidos que pudieran estar involucrados en procesos de unión a ligando, por los motivos expuestos anteriormente (sección 3.2 y 6.1). En relación con los resultados de la sección 5.1.3, se observa estos son muy distintos a las estructuras secundarias calculadas por el algoritmo de DSSP (Kabsch, 1983), lo cual puede ser consecuencia del tipo de datos empleados para entrenar a los predictores (JPred y PSIPRED) (Drozdetskiy, 2015) (Buchan, 2013) y del uso de proteínas globulares para el desarrollo del algoritmo de CFSSP (Chou, 1974a,b). Por otro lado, al tener fragmentos individuales de la BSP sus interacciones y la formación de una estructura secundaria pueden ser diferentes a lo predicho para la proteína íntegra.

El análisis de estructuras secundarias muestra que los veintidos péptidos tienen un contenido de estructura aleatoria (giros) de aproximadamente un 50%, un contenido de hélice α de entre 0 y 13% y el porcentaje restante corresponde a hojas β , siendo valores cercanos a los reportados por Wuttke (2001) para la BSP. Para los péptidos P1Ca, P2Ca, P3Ca, P4Ca, P5Ca, P5FCa y P6Ca se observa un aumento en el porcentaje de hoja β detectada por el algoritmo DSSP. Se sabe que la presencia de

iones en un ambiente acuoso puede favorecer la estabilización de la estructura secundaria de un péptido, sin embargo, los mecanismos bajo los que ocurre este fenómeno siguen siendo inciertos (Crevenna, 2012). Únicamente uno de los péptidos en los que se observa este comportamiento se encuentra fosforilado (P5FCa), interesantemente este péptido cuenta con un aminoácido (S-133) que ha sido reportado como “crucial” para el proceso de nucleación de HA (sección 3.2.2), motivo por el cual este péptido cuenta con un indicador de importancia para la selección de péptidos en futuros proyectos.

La influencia sobre la presencia de hélices α es menor a la observada en hojas β , pues los cambios en contenido de éstas son de 1%-2%. A excepción del péptido P3, el cual tiene un contenido de hélice α del 13%. Este valor disminuye a 0% en presencia de cualquier modificación al péptido (péptidos P3F, P3Ca y P3FCa), que por lo observado anteriormente puede ser consecuencia de la elongación de la cadena peptídica.

El análisis de puentes de hidrógeno indica una alta capacidad por parte de los aminoácidos fosforilados para participar en la formación de estos, misma que se ve reducida, y en algunos casos perdida, en presencia de iones Ca^{2+} pues al interactuar de forma directa, se forma un grupo de mayor volumen cuyo impedimento estérico no permite la formación de puentes.

Los análisis de las trayectorias nos dan información sobre el comportamiento y posición de los iones Ca^{2+} alrededor de los péptidos. Los iones Ca^{2+} presentan una afinidad para agruparse en presencia de una alta carga negativa. Esto es observado durante las simulaciones de DM alrededor de los aminoácidos 48-83 y 132-142, y se incrementa en presencia de las fosforilaciones. Estas secuencias contienen dos octámeros de glutamato. Las secuencias anteriores se han estudiado para tratar de entender los mecanismos de nucleación mediante simulaciones de DM (Yang, 2010); sin embargo, en dicho estudio se construye el péptido asumiendo una conformación estructural inicial de hélice α . En nuestras simulaciones de DM, se observa que estas secuencias “reclutan” iones calcio en los alrededores del péptido,

lo cual es un indicador de importancia para la selección de péptidos para estudios futuros.

Los péptidos P1Ca y P6Ca, a diferencia de los otros péptidos en presencia de calcio, no muestran interacciones directas con este ion. A pesar de tener, ambos péptidos, una carga negativa.

Del análisis de distancias Ca^{2+} - residuo se sabe que los iones que se asocian a los grupos fosfato muestran una distancia inicial al aminoácido de entre 20 y 100 Å y una distancia mínima de entre 3 y 40 Å, dependiendo de la longitud del péptido, a la fecha no se reporta en bibliografía resultados de este estilo bajo condiciones similares,

7.0 Conclusiones

Se identificaron dos motivos de unión a Ca^{2+} correspondientes a los aminoácidos 48-83 y 132-142 de la BSP. Al mismo tiempo se identificó que la presencia de fosforilaciones provoca un aumento en la interacción con Ca^{2+} y diversos cambios estructurales.

A partir de distintos métodos bioinformáticos y con respaldo en la literatura, se seleccionaron seis péptidos con potencial para participar en el proceso de la biomineralización. Y mediante simulaciones de DM se identificaron dos potenciales dominios de unión a Ca^{2+} .

La presencia de Ca^{2+} parece estabilizar la formación de hojas β , mostrando un aumento en el contenido de esta estructura en todos los péptidos. Mientras que su presencia adjunta a las fosforilaciones propicia distintos estados estructurales posiblemente relacionados con el papel mecanístico de BSP en la biomineralización.

8.0 Perspectivas

Para comprobar los hallazgos encontrados en los péptidos de BSP se propone realizar estudios posteriores, como difracción de rayos x de bajo ángulo (SAXS), dispersión dinámica de luz, (DLS), dicroísmo circular (CD), espectroscopía de infrarrojo por transformada de Fourier (FTIR) y calorimetría de titulación isotérmica (ITC). Mediante las técnicas de SAXS y DLS se podrá comprobar el grado de compactación (R_g) de los péptidos con y sin fosforilaciones, además de evaluar el efecto de la adición de iones calcio. Utilizando al CD y FTIR en solución, se podrá comprobar los cambios en estructura secundaria en los péptidos debido al efecto de las fosforilaciones y a la adición de los iones calcio s. Mientras con el ITC, nos permitirá medir de manera directa el calor que se libera o absorbe durante la interacción entre los péptidos con o sin fosforilaciones y los iones calcio y se puede determinar simultáneamente todos los parámetros del enlace en un solo experimento. También, se encuentran propuestos distintas simulaciones de dinámica molecular de los distintos péptidos de BSP con HA y fibras de colágena para analizar sus interacciones.

9.0 Referencias

- Abraham, M.J, van der Spoel, D., Lindahl, E., Hess, B., and the GROMACS development team (2018), GROMACS User Manual version 2018, www.gromacs.org
- Addadi, L., & Weiner, S. (2014). Biomineralization: mineral formation by organisms. *Physica Scripta*, 89(9), 098003. <https://doi.org/10.1088/0031-8949/89/9/098003>
- Alberts, B. (2015). *Molecular biology of the cell* (Sixth edition). New York, NY: Garland Science, Taylor and Francis Group.
- Alder, B. J., & Wainwright, T. E. (1957). Phase Transition for a Hard Sphere System. *The Journal of Chemical Physics*, 27(5), 1208-1209. <https://doi.org/10.1063/1.1743957>
- Altschul, 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.
- Altschul, S., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17), 3389-3402. <https://doi.org/10.1093/nar/25.17.3389>
- Atkins, J., Boateng, S., Sorensen, T., & McGuffin, L. (2015). Disorder Prediction Methods, Their Applicability to Different Protein Targets and Their Usefulness for Guiding Experimental Studies. *International Journal of Molecular Sciences*, 16(8), 19040-19054. <https://doi.org/10.3390/ijms160819040>
- Ayodele, T. O. (2010). Machine Learning Overview. *En New Advances in Machine Learning* (pp. 9-18). InTech.
- Bah, A., Vernon, R. M., Siddiqui, Z., Krzeminski, M., Muhandiram, R., Zhao, C., ... Forman-Kay, J. D. (2015). Folding of an intrinsically disordered protein by phosphorylation as a regulatory switch. *Nature*, 519(7541), 106-109. <https://doi.org/10.1038/nature13999>
- Baht, G. S., Hunter, G. K., & Goldberg, H. A. (2008). Bone sialoprotein–collagen interaction promotes hydroxyapatite nucleation. *Matrix Biology*, 27(7), 600-608. <https://doi.org/10.1016/j.matbio.2008.06.004>
- Baht, G. S., O'Young, J., Borovina, A., Chen, H., Tye, C. E., Karttunen, M., ... Goldberg, H. A. (2010). Phosphorylation of Ser 136 is critical for potent bone sialoprotein-mediated nucleation of hydroxyapatite crystals. *Biochemical Journal*, 428(3), 385-395. <https://doi.org/10.1042/BJ20091864>
- Bayat, A. (2002). Science, medicine, and the future: Bioinformatics. *BMJ*, 324(7344), 1018-1022. <https://doi.org/10.1136/bmj.324.7344.1018>

Bellahcène, A., Castronovo, V., Ogbureke, K. U. E., Fisher, L. W., & Fedarko, N. S. (2008). Small Integrin-Binding Ligand N-linked Glycoproteins (SIBLINGs): Multifunctional proteins in cancer. *Nature reviews. Cancer*, 8(3), 212-226. <https://doi.org/10.1038/nrc2345>

Beniash, E. (2011). *Biomaterials—hierarchical nanocomposites: The example of bone*. Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology, 3(1), 47-69. <https://doi.org/10.1002/wnan.105>

Berendsen, H. J. C., van der Spoel, D., & van Drunen, R. (1995). GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, 91(1-3), 43-56. [https://doi.org/10.1016/0010-4655\(95\)00042-E](https://doi.org/10.1016/0010-4655(95)00042-E).

Bhagwat, M., & Aravind, L. (2007). *PSI-BLAST Tutorial*. Humana Press. Recuperado de <https://www.ncbi.nlm.nih.gov/books/NBK2590/>

Bishop, C. M., & Hinton, G. (1995). *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press.

Blom, N., Gammeltoft, S., & Brunak, S. (1999). Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *Journal of Molecular Biology*, 294(5), 1351-1362. <https://doi.org/10.1006/jmbi.1999.3310>

Blom, N., Hansen, J., Blaas, D., & Brunak, S. (1996). Cleavage site analysis in picornaviral polyproteins: discovering cellular targets by neural networks. *Protein Science : A Publication of the Protein Society*, 5(11), 2203-2216.

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(6), 1633-1649. <https://doi.org/10.1002/pmic.200300771>

Bloomer, A. C., Champness, J. N., Bricogne, G., Staden, R., & Klug, A. (1978). Protein disk of tobacco mosaic virus at 2.8 Å resolution showing the interactions within and between subunits. *Nature*, 276(5686), 362–368.

Boskey, A. L. (2007). Mineralization of Bones and Teeth. *Elements*, 3(6), 385-391. <https://doi.org/10.2113/GSELEMENTS.3.6.385>

Boskey, A. L. (2013). Bone composition: Relationship to bone fragility and antiosteoporotic drug effects. *BoneKEy Reports*, 2. <https://doi.org/10.1038/bonekey.2013.181>

Boskey, A. L., & Villarreal-Ramirez, E. (2016). Intrinsically disordered proteins and biomineralization. *Matrix Biology*, 52-54, 43-59. <https://doi.org/10.1016/j.matbio.2016.01.007>

Braissant, O., Cailleau, G., Aragno, M., & Verrecchia, E. P. (2004). Biologically induced mineralization in the tree *Milicia excelsa* (Moraceae): its causes and consequences to the environment. *Geobiology*, 2(1), 59-66. <https://doi.org/10.1111/j.1472-4677.2004.00019.x>

Buchan, D. W. A., Minneci, F., Nugent, T. C. O., Bryson, K., & Jones, D. T. (2013). Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Research*, 41(W1), W349-W357. <https://doi.org/10.1093/nar/gkt381>

Carluccio, C., Fraternali, F., Salvatore, F., Fornili, A., & Zagari, A. (2013). Structural Features of the Regulatory ACT Domain of Phenylalanine Hydroxylase. *PLoS ONE*, 8(11), e79482. <https://doi.org/10.1371/journal.pone.0079482>

Chen, Y., Bal, B. S., & Gorski, J. P. (1992). Calcium and collagen binding properties of osteopontin, bone sialoprotein, and bone acidic glycoprotein-75 from bone. *Journal of Biological Chemistry*, 267(34), 24871-24878.

Chen, Y., Feng, Y., Deveaux, J. G., Masoud, M. A., Chandra, F. S., Chen, H., Feng, L. (2019). Biomineralization Forming Process and Bio-inspired Nanomaterials for Biomedical Application: A Review. *Minerals*, 9(2), 68. <https://doi.org/10.3390/min9020068>

Chin, A. F., Topygin, D., Elam, W. A., Schrank, T. P., & Hilser, V. J. (2016). Phosphorylation Increases Persistence Length and End-to-End Distance of a Segment of Tau Protein. *Biophysical Journal*, 110(2), 362-371. <https://doi.org/10.1016/j.bpj.2015.12.013>

Chou, P. Y., & Fasman, G. D. (1974a). Conformational Parameters for Amino Acids in Helical, β -Sheet, and Random Coil Regions Calculated from Protein Data Bank. *Biochemistry*, 13(2), 211-222.

Chou, P. Y., & Fasman, G. D. (1974b). Prediction of Protein Conformation. *Biochemistry*, 13(2), 222-245.

Cieřła, J., Frączyk, T., & Rode, W. (2011). Phosphorylation of basic amino acid residues in proteins: important but easily missed. *Acta Biochimica Polonica* 58, 11

Coletta, A., Pinney, J. W., Solís, D. Y. W., Marsh, J., Pettifer, S. R., & Attwood, T. K. (2010). Low-complexity regions within protein sequences have position-dependent roles. *BMC Systems Biology*, 4(1). <https://doi.org/10.1186/1752-0509-4-43>

Collier, T. A., Nash, A., Birch, H. L., & de Leeuw, N. H. (2018). Relative orientation of collagen molecules within a fibril: A homology model for homo sapiens type I collagen. *Journal of Biomolecular Structure and Dynamics*, 37(2), 537-549. <https://doi.org/10.1080/07391102.2018.1433553>

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297. <https://doi.org/10.1007/BF00994018>

Crevenna, A. H., Naredi-Rainer, N., Lamb, D. C., Wedlich-Söldner, R., & Dzubiella, J. (2012). Effects of Hofmeister Ions on the α -Helical Structure of Proteins. *Biophysical Journal*, 102(4), 907-915. <https://doi.org/10.1016/j.bpj.2012.01.035>

Cuff, J. A., & Barton, G. J. (2000). Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins: Structure, Function, and Genetics*, 40(3), 502-511. [https://doi.org/10.1002/1097-0134\(20000815\)40:3<502::AID-PROT170>3.0.CO;2-Q](https://doi.org/10.1002/1097-0134(20000815)40:3<502::AID-PROT170>3.0.CO;2-Q)

Daura, X., Mark, A. E., & Van Gunsteren, W. F. (1998). Parametrization of aliphatic CHn united atoms of GROMOS96 force field. *Journal of Computational Chemistry*, 19(5), 535-547. [https://doi.org/10.1002/\(SICI\)1096-987X\(19980415\)19:5<535::AID-JCC6>3.0.CO;2-N](https://doi.org/10.1002/(SICI)1096-987X(19980415)19:5<535::AID-JCC6>3.0.CO;2-N)

Drozdetskiy, A., Cole, C., Procter, J., & Barton, G. J. (2015). JPred4: a protein secondary structure prediction server. *Nucleic Acids Research*, 43(W1), W389-W394. <https://doi.org/10.1093/nar/gkv332>

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.jmglm.2011.07.008>

Dyson, H. J. (2016). Making Sense of Intrinsically Disordered Proteins. *Biophysical Journal*, 110(5), 1013-1016. <https://doi.org/10.1016/j.bpj.2016.01.030>

Eddy, S. R. (2004). Where did the BLOSUM62 alignment score matrix come from? *Nature Biotechnology*, 22(8), 1035-1036. <https://doi.org/10.1038/nbt0804-1035>

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), 1792-1797. <https://doi.org/10.1093/nar/gkh340>

Estroff, L. A. (2008). Introduction: Biomineralization. *Chemical Reviews*, 108(11), 4329-4331. <https://doi.org/10.1021/cr8004789>

Fihri, A., Len, C., Varma, R. S., & Solhy, A. (2017). Hydroxyapatite: A review of syntheses, structure and applications in heterogeneous catalysis. *Coordination Chemistry Reviews*, 347, 48-76. <https://doi.org/10.1016/j.ccr.2017.06.009>

Fisher, L. W., McBride, O. W., Termine, J. D., & Young, M. F. (1990). Human bone sialoprotein. Deduced protein sequence and chromosomal localization. *Journal of Biological Chemistry*, 265(4), 2347-2351.

Fisher, L. W., Torchia, D. A., Fohr, B., Young, M. F., & Fedarko, N. S. (2001). Flexible Structures of SIBLING Proteins, Bone Sialoprotein, and Osteopontin. *Biochemical and Biophysical Research Communications*, 280(2), 460-465. <https://doi.org/10.1006/bbrc.2000.4146>

Francis, L. (2001). *Neural Networks Demystified*. The Casualty Actuarial Society Forum, Winter 2001, 253-320.

Frueh, D. P., Goodrich, A. C., Mishra, S. H., & Nichols, S. R. (2013). NMR methods for structural studies of large monomeric and multimeric proteins. *Current Opinion in Structural Biology*, 23(5), 734-739. <https://doi.org/10.1016/j.sbi.2013.06.016>

Ganss, B., Kim, R. H., & Sodek, J. (1999). Bone Sialoprotein. *Critical Reviews in Oral Biology & Medicine*, 10(1), 79-98. <https://doi.org/10.1177/10454411990100010401>

Gardner, M. ., & Dorling, S. . (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*, 32(14-15), 2627-2636. [https://doi.org/10.1016/S1352-2310\(97\)00447-0](https://doi.org/10.1016/S1352-2310(97)00447-0)

George, A., & Veis, A. (2008). Phosphorylated Proteins and Control over Apatite Nucleation, Crystal Growth, and Inhibition. *Chemical Reviews*, 108(11), 4670-4693. <https://doi.org/10.1021/cr0782729>

Gholami, R., & Fakhari, N. (2017). Support Vector Machine: Principles, Parameters, and Applications. En *Handbook of Neural Computation* (pp. 515-535). Elsevier. <https://doi.org/10.1016/B978-0-12-811318-9.00027-2>

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>

Gorby, Y. A., Beveridge, T. J., & Blakemore, R. P. (1988). Characterization of the bacterial magnetosome membrane. *Journal of Bacteriology*, 170(2), 834-841. <https://doi.org/10.1128/jb.170.2.834-841.1988>

Gordon, J. A. R., Tye, C. E., Sampaio, A. V., Underhill, T. M., Hunter, G. K., & Goldberg, H. A. (2007). Bone sialoprotein expression enhances osteoblast differentiation and matrix mineralization in vitro. *Bone*, 41(3), 462-473. <https://doi.org/10.1016/j.bone.2007.04.191>

Gorski, J. P. (2011). Biomineralization of bone: A fresh view of the roles of non-collagenous proteins. *Frontiers in Bioscience*, 16(1), 2598. <https://doi.org/10.2741/3875>

Hale, R. L., & Dougherty, D. (1988). Differences between ward's and UPGMA methods of cluster analysis: Implications for school psychology. *Journal of School Psychology*, 26(2), 121-131. [https://doi.org/10.1016/0022-4405\(88\)90014-3](https://doi.org/10.1016/0022-4405(88)90014-3)

Hauptmann, S., Dufner, H., Brickmann, J., Kast, S. M., & Berry, R. S. (2003). Potential energy function for apatites. *Physical Chemistry Chemical Physics*, 5(3), 635-639. <https://doi.org/10.1039/b208209h>

Heinegård, D., & Oldberg, A. (1989). Structure and biology of cartilage and bone matrix noncollagenous macromolecules. *The FASEB Journal*, 3(9), 2042-2051. <https://doi.org/10.1096/fasebj.3.9.2663581>

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89(22), 10915-10919.

Heyes, D. M. (1983). Molecular dynamics at constant pressure and temperature. *Chemical Physics*, 82(3), 285-301. [https://doi.org/10.1016/0301-0104\(83\)85235-5](https://doi.org/10.1016/0301-0104(83)85235-5)

Hunt, L(1984). Margaret Oakley Dayhoff 1925–1983. *Bulletin of Mathematical Biology*, 46(4), 467-472. <https://doi.org/10.1007/BF02459497>

Hunter, G. K., & Goldberg, H. A. (1993). Nucleation of hydroxyapatite by bone sialoprotein. *Proceedings of the National Academy of Sciences*, 90(18), 8562-8565. <https://doi.org/10.1073/pnas.90.18.8562>

Iakoucheva, L. M. (2004). The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Research*, 32(3), 1037-1049. <https://doi.org/10.1093/nar/gkh253>

Ige, O. O., Umoru, L. E., & Aribu, S. (2012). Natural Products: A Minefield of Biomaterials. *ISRN Materials Science*, 2012, 1-20. <https://doi.org/10.5402/2012/983062>

Ishida, T., & Kinoshita, K. (2007). PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Research*, 35(Web Server), W460-W464. <https://doi.org/10.1093/nar/gkm363>

Ji, Z., Lipton, Z. C., & Elkan, C. (2014). Differential Privacy and Machine Learning: a Survey and Review. *ArXiv:1412.7584 [Cs]*. Recuperado de <http://arxiv.org/abs/1412.7584>

Jones, D. T. (1999). Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices. *Journal of Molecular Biology*, 292(2), 195-202.

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12), 2577-2637. <https://doi.org/10.1002/bip.360221211>

Kadler, K. E., Baldock, C., Bella, J., & Boot-Handford, R. P. (2007). Collagens at a glance. *Journal of Cell Science*, 120(12), 1955-1958. <https://doi.org/10.1242/jcs.03453>

Kadler, K. E., Holmes, D. F., Trotter, J. A., & Chapman, J. A. (1996). Collagen fibril formation. *Biochemical Journal*, 316(1), 1-11. <https://doi.org/10.1042/bj3160001>

Kalka, M., Zoglowek, A., & Dobryczycki, P. (2019). Proteins in Calcium Phosphates Biomineralization. In *Calcium Phosphates - From Fundamentals to Applications* (pp. 1-18). InTech.

Kalmar, L., Homola, D., Varga, G., & Tompa, P. (2012). Structural disorder in proteins brings order to crystal growth in biomineralization. *Bone*, 51(3), 528-534. <https://doi.org/10.1016/j.bone.2012.05.009>

Katoh, K., & Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics*, 9(4), 286-298. <https://doi.org/10.1093/bib/bbn013>

Katoh, K., Kuma, K., Toh, H., & Miyata, T. (2005). MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, 33(2), 511-518. <https://doi.org/10.1093/nar/gki198>

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14), 3059-3066. <https://doi.org/10.1093/nar/gkf436>

Kerfeld, C. A., & Scott, K. M. (2011). Using BLAST to Teach “E-value-tionary” Concepts. *PLoS Biology*, 9(2). <https://doi.org/10.1371/journal.pbio.1001014>

Kim, R. H., Shapiro, H. S., Li, J. J., Wrana, J. L., & Sodek, J. (1994). Characterization of the Human Bone SialoProtein (BSP) Gene and its Promoter Sequence. *Matrix Biology*, 14(1), 31-40. [https://doi.org/10.1016/0945-053X\(94\)90027-2](https://doi.org/10.1016/0945-053X(94)90027-2)

Kohn, W. D., Kay, C. M., & Hodges, R. S. (2008). Protein destabilization by electrostatic repulsions in the two-stranded α -helical coiled-coil/leucine zipper. *Protein Science*, 4(2), 237-250. <https://doi.org/10.1002/pro.5560040210>

Koua, D., Laht, S., Kaplinski, L., Stöcklin, R., Remm, M., Favreau, P., & Lisacek, F. (2013). Position-specific scoring matrix and hidden Markov model complement each other for the prediction of conopeptide superfamilies. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1834(4), 717-724. <https://doi.org/10.1016/j.bbapap.2012.12.015>

Kreegipuu, A., Blom, N., & Brunak, S. (1999). PhosphoBase, a database of phosphorylation sites: Release 2.0. *Nucleic Acids Research*, 27(1), 237-239.

- Krieger, E., & Vriend, G. (2014). YASARA View—molecular graphics for all devices—from smartphones to workstations. *Bioinformatics*, 30(20), 2981-2982. <https://doi.org/10.1093/bioinformatics/btu426>
- Kröger, R. (2015). Ion binding and nucleation: Biomineralization. *Nature Materials*, 14(4), 369-370. <https://doi.org/10.1038/nmat4256>
- Kruzic, J. J., & Ritchie, R. O. (2008). Fatigue of mineralized tissues: Cortical bone and dentin. *Journal of the Mechanical Behavior of Biomedical Materials*, 1(1), 3-17. <https://doi.org/10.1016/j.jmbbm.2007.04.002>
- Kulkarni, P., Jolly, M. K., Jia, D., Mooney, S. M., Bhargava, A., Kagohara, L. T., ... Orban, J. (2017). Phosphorylation-induced conformational dynamics in an intrinsically disordered protein and potential role in phenotypic heterogeneity. *Proceedings of the National Academy of Sciences*, 114(13), E2644-E2653. <https://doi.org/10.1073/pnas.1700082114>
- Kumar, T. A. (2013). CFSSP: Chou and Fasman Secondary Structure Prediction Server. *Wide Spectrum*, 1(9), 15-19
- Kumar, T. A. (2013). CFSSP: Chou and Fasman Secondary Structure Prediction Server. *Wide Spectrum*, 1(9), 15-19.
- Lai, K., Twine, N., O'Brien, A., Guo, Y., & Bauer, D. (2019). Artificial Intelligence and Machine Learning in Bioinformatics. En *Encyclopedia of Bioinformatics and Computational Biology* (pp. 272-286). Elsevier. <https://doi.org/10.1016/B978-0-12-809633-8.20325-7>
- Li, X, Romero, P, Rani M, Dunker AK, and Obradovic Z. (1999) Predicting protein disorder for N-, C-, and internal regions, *Genome Informatics*, 10, 30-40. <https://doi.org/10.11234/gi1990.10.30>
- Lin, K., & Chang, J. (2015). Structure and properties of hydroxyapatite for biomedical applications. En *Hydroxyapatite (Hap) for Biomedical Applications* (pp. 3-19). <https://doi.org/10.1016/B978-1-78242-033-0.00001-8>
- Lobanov, M. Y., Bogatyreva, N. S., & Galzitskaya, O. V. (2008). Radius of gyration as an indicator of protein structure compactness. *Molecular Biology*, 42(4), 623-628. <https://doi.org/10.1134/S0026893308040195>
- Lowenstam, H. A., & Weiner, S. (1989). *On Biomineralization*. New York, NY: Oxford University Press.
- Mackay, D. H. J., Cross, A. J., & Hagler, A. T. (1989). The Role of Energy Minimization in Simulation Strategies of Biomolecular Systems. En G. D. Fasman (Ed.), *Prediction of*

Protein Structure and the Principles of Protein Conformation (pp. 317-358).
https://doi.org/10.1007/978-1-4613-1571-1_7

Mackerell, A. D. (2004). Empirical force fields for biological macromolecules: Overview and issues. *Journal of Computational Chemistry*, 25(13), 1584-1604.
<https://doi.org/10.1002/jcc.20082>

Maglogiannis, I. G. (2007). *Emerging Artificial Intelligence Applications in Computer Engineering: Real World AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies*. IOS Press.

Maguid, S., Fernández-Alberti, S., Parisi, G., & Echave, J. (2006). Evolutionary Conservation of Protein Backbone Flexibility. *Journal of Molecular Evolution*, 63(4), 448-457. <https://doi.org/10.1007/s00239-005-0209-x>

McGuffin, L. J., Bryson, K., & Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics*, 16(4), 404-405. <https://doi.org/10.1093/bioinformatics/16.4.404>

McKee, M. D., & Cole, W. G. (2012). Chapter 2—Bone Matrix and Mineralization. *En Pediatric Bone* (2nd ed., pp. 9-37). NY: Elsevier/Academic Press.

McKee, M. D., Addison, W. N., & Kaartinen, M. T. (2005). Hierarchies of Extracellular Matrix and Mineral Organization in Bone of the Craniofacial Complex and Skeleton. *Cells Tissues Organs*, 181(3-4), 176-188. <https://doi.org/10.1159/000091379>

McLachlan, A. D. (1972). Protein Structure and Function. *Annual Review of Physical Chemistry*, 23(1), 165–192.

Miller, A., & Parker, S. B. (1984). Collagen: The Organic Matrix of Bone [and Discussion]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 304(1121), 455-477. <https://doi.org/10.1098/rstb.1984.0040>

Mount, W. David (2001). Database Searching for Similar Sequences. *En Bioinformatics: Sequence and Genome Analysis* (pp 281-336). Cold Spring Harbor Laboratory Press

Nancollas, G. H. (1977). The mechanism of biological mineralization. *Journal of Crystal Growth*, 42, 185-193. [https://doi.org/10.1016/0022-0248\(77\)90194-4](https://doi.org/10.1016/0022-0248(77)90194-4)

Nuin, P. A., Wang, Z., & Tillier, E. R. (2006). The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics*, 7, 471. <https://doi.org/10.1186/1471-2105-7-471>

Pace, C. N., Fu, H., Lee Fryar, K., Landua, J., Trevino, S. R., Schell, D., ... Grimsley, G. R. (2014). Contribution of hydrogen bonds to protein stability: Hydrogen Bonds and Protein Stability. *Protein Science*, 23(5), 652-661. <https://doi.org/10.1002/pro.2449>

Pasteris, J. D., Wopenka, B., & Valsami-Jones, E. (2008). Bone and Tooth Mineralization: Why Apatite? *Elements*, 4(2), 97-104. <https://doi.org/10.2113/GSELEMENTS.4.2.97>

Perovic, V., Sumonja, N., Marsh, L. A., Radovanovic, S., Vukicevic, M., Roberts, S. G. E., & Veljkovic, N. (2018). IDPpi: Protein-Protein Interaction Analyses of Human Intrinsically Disordered Proteins. *Scientific Reports*, 8(1), 10563. <https://doi.org/10.1038/s41598-018-28815-x>

Ponting, C. P. (2017). Biological function in the twilight zone of sequence conservation. *BMC Biology*, 15(1), 71. <https://doi.org/10.1186/s12915-017-0411-5>

Posner A.S., Blumenthal N.C., Betts F. (1984) Chemistry and Structure of Precipitated Hydroxyapatites. En: Nriagu J.O., Moore P.B. (eds) Phosphate Minerals. Springer, Berlin, Heidelberg

Prevelige, P., & Fasman, G. D. (1989). Chou-Fasman Prediction of the Secondary Structure of Proteins. En G. D. Fasman (Ed.), *Prediction of Protein Structure and the Principles of Protein Conformation* (pp. 391-416). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4613-1571-1_9

Radivojac, P., Obradovic, Z., Smith, D. K., Zhu, G., Vucetic, S., Brown, C. J., ... Dunker, A. K. (2004). Protein flexibility and intrinsic disorder. *Protein Science: A Publication of the Protein Society*, 13(1), 71-80. <https://doi.org/10.1110/ps.03128904>

Reber, R., & Perrig, W. (2001). Perception without Awareness, *Psychology of*. En *International Encyclopedia of the Social & Behavioral Sciences* (pp.11215-11218). Elsevier. <https://doi.org/10.1016/B0-08-043076-7/01461-3>

Romero P, Obradovic Z, and Dunker AK. (1997) Sequence data analysis for long disordered regions prediction in the calcineurin family, *Genome Informatics*, 8, 110-124

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(1), 38-48. [https://doi.org/10.1002/1097-0134\(20010101\)42:1<38::AID-PROT50>3.0.CO;2-3](https://doi.org/10.1002/1097-0134(20010101)42:1<38::AID-PROT50>3.0.CO;2-3)

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386-408. <https://doi.org/10.1037/h0042519>

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986a). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1. En David E. Rumelhart, J. L.

McClelland, & C. PDP Research Group (Eds.) (pp. 318–362). Cambridge, MA, USA: MIT Press.

Rumelhart, D. E., Hintont, G. E., & Williams, R. J. (1986b). Learning representations by back-propagating errors. *Nature*, 323(6099), 533-536. <https://doi.org/doi:10.1038/323533a0>.

Salih, E. (2003). In Vivo and In Vitro Phosphorylation Regions of Bone Sialoprotein. *Connective Tissue Research*, 44(1), 223-229. <https://doi.org/10.1080/03008200390181708>

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 3(3), 210-229.

Schaffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I., Altschul, S. (2001). Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research*, 29(14), 2994-3005. <https://doi.org/10.1093/nar/29.14.2994>

Scott, W. R. P., Hünenberger, P. H., Tironi, I. G., Mark, A. E., Billeter, S. R., Fennen, J., van Gunsteren, W. F. (1999). The GROMOS Biomolecular Simulation Program Package. *The Journal of Physical Chemistry A*, 103(19), 3596-3607. <https://doi.org/10.1021/jp984217f>

Shoemaker, B. A., Portman, J. J., & Wolynes, P. G. (2000). Speeding molecular recognition by using the folding funnel: The fly-casting mechanism. *Proceedings of the National Academy of Sciences*, 97(16), 8868-8873. <https://doi.org/10.1073/pnas.160259697>

Shoulders, M. D., & Raines, R. T. (2009). Collagen Structure and Stability. *Annual Review of Biochemistry*, 78(1), 929-958. <https://doi.org/10.1146/annurev.biochem.77.032207.120833>

Simkiss, K., & Wilbur, K. M. (2012). *Biom mineralization* (1.^a ed.). New York. Elsevier.

Sitbon, E., & Pietrokovski, S. (2007). Occurrence of protein structure elements in conserved sequence regions. *BMC Structural Biology*, 7(1), 1-15. <https://doi.org/10.1186/1472-6807-7-3>

Sneha, P., & George Priya Doss, C. (2016). Molecular Dynamics. In *Advances in Protein Chemistry and Structural Biology* (Vol. 102, pp. 181-224). Elsevier. <https://doi.org/10.1016/bs.apcsb.2015.09.004>

Steeg, P. S., Palmieri, D., Ouatas, T., & Salerno, M. (2003). Histidine kinases and histidine phosphorylated proteins in mammalian cell biology, signal transduction and cancer. *Cancer Letters*, 190(1), 1-12. [https://doi.org/10.1016/S0304-3835\(02\)00499-8](https://doi.org/10.1016/S0304-3835(02)00499-8)

Stubbs, J. T., Mintz, K. P., Eanes, E. D., Torchia, D. A., & Fisher, L. W. (1997). Characterization of Native and Recombinant Bone Sialoprotein: Delineation of the Mineral-Binding and Cell Adhesion Domains and Structural Analysis of the RGD Domain. *Journal of Bone and Mineral Research*, 12(8), 1210-1222. <https://doi.org/10.1359/jbmr.1997.12.8.1210>

Suárez, E. J. C. (2016). Tutorial sobre Máquinas de Vectores Soporte (SVM), Dpto. de Inteligencia Artificial, ETS de Ingeniería Informática, Universidad Nacional de Educación a Distancia (UNED), Madrid. Pp

Sun, X., Wei, J., Lyu, J., Bian, T., Liu, Z., Huang, J., ... Zhong, Z. (2019). Bone-targeting drug delivery system of biomineral-binding liposomes loaded with icariin enhances the treatment for osteoporosis. *Journal of Nanobiotechnology*, 17(1), 10. <https://doi.org/10.1186/s12951-019-0447-5>

Termine, J. D., Belcourt, A., Conn, K., & Kleinman, S. H. (1981). Mineral and Collagen-binding Proteins of Fetal Calf Bone. *Journal of Biological Chemistry*, 256, 10403-10408.

Tye, C. E., Hunter, G. K., & Goldberg, H. A. (2005). Identification of the Type I Collagen-binding Domain of Bone Sialoprotein and Characterization of the Mechanism of Interaction. *Journal of Biological Chemistry*, 280(14), 13487-13492. <https://doi.org/10.1074/jbc.M408923200>

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(10), 7949-7955. <https://doi.org/10.1074/jbc.M211915200>

Uversky, V. N., & Dunker, A. K. (2010). Understanding protein non-folding. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1804(6), 1231-1264. <https://doi.org/10.1016/j.bbapap.2010.01.017>

Valdar, W. S. J. (2002). Scoring residue conservation. *Proteins: Structure, Function, and Genetics*, 48(2), 227-241. <https://doi.org/10.1002/prot.10146>

van der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., ... Babu, M. M. (2014). Classification of Intrinsically Disordered Regions and Proteins. *Chemical Reviews*, 114(13), 6589-6631. <https://doi.org/10.1021/cr400525m>

Veis, A., Sfeir, C., & Chou Bing Wu. (1997). Phosphorylation of the Proteins of the Extracellular Matrix of Mineralized Tissues By Casein Kinase-Like Activity. *Critical Reviews in Oral Biology & Medicine*, 8(4), 360-379. <https://doi.org/10.1177/10454411970080040101>

Verlet, L. (1967). Computer «Experiments» on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review*, 159(1), 98-103. <https://doi.org/10.1103/PhysRev.159.98>

Villarreal-Ramirez, 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-Ramirez, E., Garduño-Juarez, R., Gericke, A., & Boskey, A. (2014). The role of phosphorylation in dentin phosphoprotein peptide absorption to hydroxyapatite surfaces: A molecular dynamics study. *Connective Tissue Research*, 55(sup1), 134-137. <https://doi.org/10.3109/03008207.2014.923870>

Vincent, K., & Durrant, M. C. (2013). A structural and functional model for human bone sialoprotein. *Journal of Molecular Graphics and Modelling*, 39, 108-117. <https://doi.org/10.1016/j.jmgm.2012.10.007>

Vlachakis, D., Bencurova, E., Papangelopoulos, N., & Kossida, S. (2014). Current State-of-the-Art Molecular Dynamics Methods and Applications. En *Advances in Protein Chemistry and Structural Biology* (Vol. 94, pp. 269-313). Elsevier. <https://doi.org/10.1016/B978-0-12-800168-4.00007-X>.

Vojisavljevic, V., & Pirogova, E. (2016). Prediction of intrinsically disordered regions in proteins using signal processing methods: application to heat-shock proteins. *Medical & Biological Engineering & Computing*, 54(12), 1831-1844. <https://doi.org/10.1007/s11517-016-1477-x>

Vullo, A., Bortolami, O., Pollastri, G., & Tosatto, S. C. E. (2006). Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. *Nucleic Acids Research*, 34(Web Server), W164-W168. <https://doi.org/10.1093/nar/gkl166>

Wassenaar, T. A. (2006). Molecular dynamics of sense and sensibility in processing and analysis of data. Groeningen Biomolecular Sciences and Biotechnology Institute, Países Bajos.

Wojtas, M., Dobryszcki, P., & Oyhar, A. (2012). Intrinsically Disordered Proteins in Biomineralization. En J. Seto (Ed.), *Advanced Topics in Biomineralization*. <https://doi.org/10.5772/31121>

Wright, P. E., & Dyson, H. J. (2015). Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology*, 16(1), 18-29. <https://doi.org/10.1038/nrm3920>

Wu, C. H. (1997). Artificial neural networks for molecular sequence analysis. *Computers & Chemistry*, 21(4), 237-256. [https://doi.org/10.1016/S0097-8485\(96\)00038-1](https://doi.org/10.1016/S0097-8485(96)00038-1)

Wuttke, M., Müller, S., Nitsche, D. P., Paulsson, M., Hanisch, F.-G., & Maurer, P. (2001). Structural Characterization of Human Recombinant and Bone-derived Bone Sialoprotein: FUNCTIONAL IMPLICATIONS FOR CELL ATTACHMENT AND HYDROXYAPATITE BINDING. *Journal of Biological Chemistry*, 276(39), 36839-36848. <https://doi.org/10.1074/jbc.M105689200>

Xu, Q., Dai, H., Zhao, T., & Wei, D. (2015). Introduction to Structural Bioinformatics. En D. Wei, Q. Xu, T. Zhao, & H. Dai (Eds.), *Advance in Structural Bioinformatics* (Vol. 827, pp. 1-7). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-017-9245-5_1

Yang, Y., Cui, Q., & Sahai, N. (2010). How Does Bone Sialoprotein Promote the Nucleation of Hydroxyapatite? A Molecular Dynamics Study Using Model Peptides of Different Conformations. *Langmuir*, 26(12), 9848-9859. <https://doi.org/10.1021/la100192z>

Young, M. F., Kerr, J. M., Ibaraki, K., Heegaard, A.-M., & Robey, P. G. (1992). Structure, Expression, and Regulation of the Major Noncollagenous Matrix Proteins of Bone. *Clinical Orthopaedics and Related Research* (281), 275-294. doi:10.1097/00003086-199208000-00042

Yuan, Z., Bailey, T. L., & Teasdale, R. D. (2005). Prediction of protein B-factor profiles. *Proteins: Structure, Function, and Bioinformatics*, 58(4), 905-912. <https://doi.org/10.1002/prot.20375>

Zaia, J., Boynton, R., Heinegård, D., & Barry, F. (2001). Posttranslational Modifications to Human Bone Sialoprotein Determined by Mass Spectrometry. *Biochemistry*, 40(43), 12983-12991. <https://doi.org/10.1021/bi010887r>

10.0 Anexos

Anexo 1: Nomenclatura de aminoácidos

A continuación, se muestra una tabla que contiene los códigos empleados a lo largo del presente trabajo, la cual contiene los 20 aminoácidos convencionales y la versión fosforilada de tres de ellos.

Aminoácido	Código tres letras	Código de una letra	Aminoácido	Código de tres letras	Código de una letra
Alanina	Ala	A	Metionina	Met	M
Arginina	Arg	R	Fenilalanina	Phe	F
Asparagina	Asn	N	Prolina	Pro	P
Ácido Aspártico	Asp	D	Serina	Ser	S
Cisteína	Cys	C	Treonina	Thr	T
Glutamina	Gln	Q	Triptófano	Trp	W
Ácido Glutámico	Glu	E	Tirosina	Tyr	Y
Glicina	Gly	G	Valina	Val	V
Histidina	His	H	Fosfo-serina	Sep	Sp
Isoleucina	Ile	I	Fosfo-tirosina	Ptr	Yp
Leucina	Leu	L	Fosfo-treonina	Tpo	Tp
Lisina	Lys	K			

Anexo 2: Secuencias de BSP

Humano

FSMKNLHRRVKIEDSEENG VFKYRPRYYL YKHA YFYPHLKRFPVQGSSDSSEEN
GDDSSEEEEEEEETSNEGENNEELNEDEDSEAENTTLSATTLYGEDATPGTGY
TGLAAIQLPKKAGDITNKATKEKESDEEEEEEEEGNENESEAEVDENEQGINGT
STNSTEAENGNSSGEDNGEEGEEESVTGANAEGTTETGRQKGKTSKTTTSPN
GGFEPTTPPVYRTTSPFPGKTTTVE YEGE YE YTGANDYDNGYE IYSENGEPR
GDNYRAYEDEYSYFKGQGYDGYDGQNYHHQ

Chimpancé

FSMKNLHRRVKIEDSEENG VFKYRPRYYL YKHA YFYPHLKRFPVQGSSDSSEEN
GDDSSEEEEEEEETSNEGENNEESNEDEDSEAENTTLSATTLYGEDTTPGTGY
TRLAAIHLPKKAGDITNKATKEKESDEEEEEEEENENESEAEVDENEQGINGTS
TNSTEAENGNSSGGDNGEEGEEESVTGANAEGTTETGGQKGKSSKTTTSPNG
GFEP TTPPVYRTTSPFPGKTTTVE YEGE YE YTGANEYDNGYE IYSENGEPRG
DNYRAYEDEYSYFKGQGYDGYDGQNYHHQ

Macaco Rhesus

FSMKNLHRRVKIEDSEENG VFKYRPRYYL YKHA YFYPLKRFPVQGSSDSSEEN
GDDSSEEEEXXETSNEGENNEESNEDEDSEAENTTLSATTLYGEDATPGTWY
TGLAAIQLPKKAGDVTNKATKKEESDEEEEEEEENENESEAEVDENEQGINGT
STNSTEAENGNSSGGDNGEEGEEESVTGASAEGTTATGGQKGKSSKTTTSP
NGGFEP TTPPVYRTTSLPFGKTTTVE YEGE YE YTGANEYDNGYEVYSENGEP
RGDNYRAYEDEYSYFKGQGYDSYDGQNYHHQ

Gorila

FSMKNLHRRVKIEDSEENG VFKYRPRYYL YKHA YFYPHLKRFPVQVNIEILEHNYN
GKTAEKEKFMTCTKKETSNEGENNEESNEDEDSEAENTTLSATTLYGEDATP
GTGYAGLAAIQLPKKAGDITNKATKEKESDEEEEEEEENENESEAEVDENEQ
INGTSTNSTEAENGNSSGGDNGEEGEEESVTGANAEGTTETGGQKGKTSKTT

TSPNGGFEP TTPPQVYRTTSP PFGKITTVE YEGE YE YTGANE YDNGYE IYSENG
EPRGDNYRAYEDEYS YFKGQGYDGYDGQNYHHQ

Tupaya

FSMKNLHRRAKLEDSEENGVFKYRPRYYLYKHA YFY PPLRRFPVQSSDSSEENG
DGDSSEEEEEEGEEASNEEENNEDSNGNEDEDSEAENSTLSTTTLGYGEEITPGT
GDTGLAAIQLPKEAGDKRNKATKQKESDEEEEEEEENENEENAEVDENEQGI
NGTSTNSTEVDNGNGSSGGDNGEEEGEEESVTGAVAEGTTVTRGQDNGGSKT
TTSSNGGLVPTTPPQIYGTTSVSYEE YEHTGTNHYDNE YEVYENENGEPRGDNY
RAYEDEYSYKGGQGYGGYGDQDYYYHQ

Ballena enana

LSMKNLHRRAKLEDSEENGVFKYRARYLYKHA YFY PPLKRFPVQSSSDLSEEN
GNDDSSEEEEEETSNEEENNEENEDSDGNEDEESEAEENTTL SITTLGYNKATKE
KESDEEEEEEEEGEETTPGTGNIGLAAIQLPTKAGDIGKKVTKEEESDEEEEEENE
NEAEADDNEQGINGTSTNGTEIDNGHGSSGGDNGEEEGEEGVTEANAEGTTVA
GGQDNGGSKATTSPNGGVGPTTPPQEIDGTTPPPFGETTTPGYEGEYEQTGTN
EYDNGYEIYENENGEPRGDNYRAYEDEYSYKGRSYDGYDGQDYYHQ

Delfín nariz de botella

LSMKNLHRRGKLEDSEENGVKYRPRYYLYKHA YFY PPLKRFPVQSSSDLSEEN
ENGESSEEEEEKETSNEEENNEENEDSEGNEDEESEAEENTTL SITTLGYGEETAP
GTGYIGLAAIQLPTQAGDIGKAATEEEESDEEGGEENENEAEVDDNEQGVNGTS
ANSTEIDNGHGSSGGDNGEEEGEEDVTEANAEGTTVAGGQDNGGSKATTSPNG
GFEP TAPPQEIYGTTPPPSGETTTPGYEGEYEQTGTNE YDNGYEIYENENGEPR
GDNYRAYEDEYSYKGRSYDGYDGQDYYYHQ

Caballo

LSMKNLHRRAKLEDSEENGVKYRPQYSLYKHA YFY PPLKRFPVQTSSDSSEEN
GDGDSSEEEEEETSNEEGNNGENEDSEGNEDEESDAENTTLSTITPGYGEEP
TPGMGNIGLAAIQLPKKAGDKGNKGTKKEEESDEEEEEEEEDENEENAEVDENGQ
GINGTSSNSTEAENHGSSGGDNGEEEGEEANAEGTAVPAEQDNAGSKTTTS

PNGGFAPTPPQEVYRTTTTA YEGE YEQTGTNE YDNGYEVYENENGEPRGDNY
RAYEDEYSYKGRSYDGYDGQDYYHQ

Murciélago

FSMKNFHRRAKLEDSEENG VFKYRPRYYLYKHA YFYPLKRFPVQNSDSSEENG
DGDSSEEEEEEGEEENSNEE ANNEESNEDEESEAENSTLSATTPGYGEETTPGTGN
VGLAAIQLPKKAGDVGKKATKKEESEEEEEEEEEENENEENAEVDENGQGINGT
STNSTDVENGNSSGGDN GEEEEEGEEENAEGTTVAGEQDNGGFEP TTPPQEV
YGTTPPPQGGTTTTTE YEGE YEQTGTNE YDNGYEVYENENAEPRGDNYRAYEDE
YSYKGRGYDSYDGQNYYYHQ

Myotis de David

FSMKNFHRRAKLEDSEENAVFKYRPRYYLYKHA YFYPLKRFAVQNSDSSEENG
DGDSSEEEEEENENEENAEVDENGQGVNGTSTNSTDAENGNSSGGDN GE
EGEEEGEEENAEGTTVAGEQDNGGSEPTTPPQEA YGTTPPPQGGTTTTTE YEEE
YEQTGTNE YDNGYEVYENENAEPRGDNYRAYEDE YSYKGRGYDSYDGQNY
YHQ

Colugo

FSMKNLHRRVKVEDSEENG VFKYRPRYSLYKHF YFYPLKRFPVQSSSDSSEEN
GDGDSSEEEEEEEETSNEE ENNENS DENEDEESEAENTTLSTATPGYGEITPGT
GIIQLAAIHLPKKAGDIGKKA AKQEE SDEEEEEEEEEEDENEENAEVDENEQGING
TSTNSTEVDNGHGSSGGDN GEEDGE GEGVTDANGEGTTAPGEWDNGSSETTT
SPNGGFQPTTPPLEGYGTTSPPF GKTTTAE YEGE YEQTGTNE YDNGYEVYD HEN
GEPRGDNYRAYEDEYSYKGRGYDGYHGQDYYHQ

Toro/Vaca

LSMKNLNRRAKLEDSEENG VFKYRPQYYVYKHGYFY PALKRFAVQSSSDSSEEN
GNGDSSEEEEEEEETSNEE GNNGGNEDSDENEDEESEAENTTLSTTTLYGEIT
PGTGDIGLAAMLPRKAGATGKKATKEDESDEEEEEEEEEENAEVDDNEQGING
TSSNSTEVDNGHGSSGGDN GEEDGE EESVTEANTEGITVAGETTTSPNGGFKPT

TPHQEVYGTTPPPFGKITTPGEYEQTGTNEYDNGYEIYSENGDPRGDNYRAYE
DEYSYKGRGYDSYDGQDYSHQ

Elefante africano

FSMKNLHRRAKLEDSEENRVFKFRPQYLYKHAYFYPPQKRFPVLSNSDSSEEN
GDGYSSEEEEEESSNEGENNEENEESENENEESEEAENTTLSTATPGYGEET
TPGTGSLGLAAIQLPKKAGHTGNKVTKTEESDEEEEEENENEENEAEDDENEQ
GINGTSTNSTDTE NGNGSNGGDHGE EEEGEEESVTEANAEEATVTGEQGNNGLE
TTASPNAGFQPTAPPQDIYGTTPPFGEITTTGYEYEQTGTNEYDNGYEVYENEN
GEPRGDNYRAYEDEYSYKGRSYDGYDGQDYNNH

Manatí

FSMKHLHRRAKLEDSEENRVFKFRPQYLYKHAYFYPPPLKRFPVLSNSDSSEEN
GDGNSSDEEEEAESSNEEENNEENEESENENEDEESEEAENTTPSTATPGYGETT
PGTGSIGLAAIQLPEKAGHIGKKVTKKEESDEEEEEENENEENEAEDDENEQGIN
GTSTNSTDAENGNSSGGDNGE EEEGGEESVTEANAETATMTGEQDNGGVETTT
SPNDGFEP TTPPQEIYGT TASPFG EITTTGYE GEYEQTGTNEYDNGY EIYENENG
EPRGDNYRAYEDEYSYKQSYDGYGQDYNNQ

Perro

FSMKNLHRRAKLEDSEENG VFKYRPRYLYKHSYFYPPPLKRFPVQSSSDSSEED
GDGDSSEEEEEEEETSNEEENNEENANSDENED ESDAENSTLSATPGYGEETP
GTGYIGLAAIQLPKKAGDIRHKATKEEESDEEEEEDEENEENEAEDDENGQGIN
TSSNSTEAENGNSSAGDN GEGEEESVTEAHSEGTT EAGKQNNGGSKTTLSPD
GGFEPTTPPELYGTTTRPSGEATPNGYEEYEQTGTNEYDNGYEVYSENGEP
RGDNYRAYEDEYSYKGHSYDSYDGQDYHHQ

Gato

FSTKNLHRRAKLEDSEENG VFKYRPRYLYKNA YFYPPPLKRFPVQSSSDSSEEN
GDGDSSEEEEEEEASNEEENNEENANSDENED ESEEAENTTL SATPGYGEETT
PGTGNIGLAAIQLPKKAGDTKKATREESDEEEEEDEENEETEEVDENEQGINST

STSTNSTEVENGNGSSDGNTGEEGEEESVTEVNAEGTTMAGRQDNGGSTTTTS
PNGGFQPTTTPPEFYGTTIRPSREATTTGYEEEEYEQTGTNEYDNGYEVYENENG
EPRGDNYRAYEDEYSYYKGRSYDSYDGQDYYYHQ

Jabali

FSMKNFHRRAKLEDPEENGVFKYRPRYYLYKHA YFYPLKRFPVQSSSDSSEEN
GNGDSSEEEEEEEENSNEEENNEENEDSDGNEDEDSEAENITLSTTTLGYGSDV
TPGTASIGLAALQLPKKAGDIGKKSAAKEEESDEDEEEEEENEENEAEVDDNEQGT
NGTSTNSTEVDSGNGHSGGDNGEEGDQESVTEAQGTTVAGEQDNGGAKTTTS
PNGGLEPTPPPQDISGTTLPPSGKTTTPE YEGE YEQTGAHE YDNGYE IYSENGE
PRGDSYRAYEDEYSYYKGRSYNSYGGHDYY

Conejo

FSMKNLHRRVKLQDSEENGVFKFRPQYYLYKHA YFYPLKRFPVPSSSDSSEENG
EGDSSDEEEEGEETSNEEENNEDSNGNEDEESEAEENSTLSFTTPDYGEITRGTG
YIGLAGIKLPKAGAIENKATEEEESDEEEEEEEENEENEAEVDENEQSINGTS
SNSTEANGEAGEGESVTATDAAGTATRGEWISRAAETTTAPSGGFEP TTPPE
AYGTTSPFGKTTAAEYGGE YEQTGNHGYDDGYE IYENENG DPRGDNYRAYED
EYSYYKGRGYEGYEGQDYYYHQ

Topo de nariz estrellada.

FSMRNLYRRAKLQDSEENGVFKYWPYSL YKHA YFYPLKRFPVQSSSDSSEEN
GDGDSSEEEEEEEETSNEEENNEENENSNGNEDEESEADNTTLSATTPGYGEEST
PALVTLGLAAIQLPKKAGPTGNKSTKKEESDEEEEEEEDEENENEENEAEVDENEQ
GINGTSTNSTEVDNGNGSSGGDNGEEGEEESATESNAEATTVAGRQDKAITKTP
SPNGGFEATTPPEVYGTTPPLGETTTPGYEE YEQTGTSE YGYEVYENENGE
RGDNYRAYEDEYSYYKGQGYGGYAGQDYY

Musaraña elefante

FSMKNLHRRAKIEDSEENGIFKFRPRYYLYKHA YFYPHLKRFPVQANSDDSSEEDG
DGNSSDEEEEEEEVRNESSNEEEEEENEDEESEANSTVSTTTFGYGEEPTPGTGGI

GLAAIQLPKKAGDAGSKATKQEESEEEEEEEEEENEENAEVDGNEQGINGTS
TNSTDVENGNSSDGDNGEEEGGEENVTKASVEGTTAPAEQETTPYANGRFEL
TTTSFGESTTTGYEDEYEQTGTNEYNNGYETYDNENTEPRGDTYRAYEDEYSYF
KGQSYDGYDSHNYYYNQ

Rata gris

FSMKNFHRIKAEDSEENG VFKYRPRYFLYKHA YFYPLKRFPVQGGSDSSEEN
GDGDSSEEEGEEEEETSNEEENNEDESEGNEDEQEAENATLSGVTASYGVETTA
DAGKLELAAALQLPKKAGDAEGKAPKMKESDEEEEEEEEEEEEEENENEEAEVDENE
QVVNGTSTNSTEVDGGNGPSGGDNGEEAEEASVTEAGAEGTTAGARELTSYGT
TTAVLLNGFQQTPPEAYGTTSPARKSSTVEYGEEYEQIGNEYNTAYETYDEN
NGEPRGDTYRAYEDEYSYKGGHGYEGYEGQDYYYHQ

Ratón

FSMKNFHRIKAEDSEENG VFKYRPRYFLYKHA YFYPLKRFPVQGGSDSSEEN
GDGDSSEEEGEEEEETSNEEENNEDESEGNEDEQEAENATLSTLSGVTASYGAET
TPQAQTFELAAALQLPKKAGDAESRAPKVKESDEEEEEEEEEEEEEENENEEAEVDEN
ELAVNGTSTNSTEVDGGNGSSGGDNGEEAEEASVTEAGAEGTTGGRELTSV
GTQTAVLLNGFQQTPPEAYGTTSPPIRKSSTVEYGGEEYEQTGNEYNNEYEVY
DNENGEPRGDTYRAYEDEYSYKGGHGYEGYEGQNYYYHQ

Armadillo de nueve bandas

FSMKNLYRRAKVEDSEENG VFKYRPRYLYKHA YFYPLKRFPVQSNSDSSEEN
GNGDSSEEEEEAGNIGSKEEESDEEEEEENETEENAEVDENEQGVNGTSTDST
DVENGNSSGGDNGEGETEDNPEGTTTTIGQDNGRFKTTASPNGGFEPTTPSH
KIYRTTTPSYRKTTEYEEYEQTGTDAYDNGYEVYENEKGVPRGDSYLGYEDEY
YYNKGHDYGNQQYYKR

Ornitorrinco

FSVKNLHRAKVEESEENAVFKHRPRYIYKYALYPPQKRILSQETSNEEENEN
EEEDENSEGEESTTVSATTPLYGEE TTPGRGDNGLGAIKVPKKAETRKSTKKEES

DEEEEEEEENENESEAEVEENEQGINGTSTNSTEDSGNDGGENGGEAAEEE
EESVTEANQEENGGQETTTFVDSVQGATTPSNGAYEVTTTPPTTSNGYEETVTNE
YDSGYEVYENENAEPRGDNFRTYEDEYSNYKGRGYDGYDGRDYYYNQ

Colicorto gris

FSAKHLYRRAKLEDSEENGVFNHRHRYL NKYSYFYPLK RIPVQHNGDSSEEG
DGDSSEEEEEETSNEEESETNGSEDEEEEGGESTTLSATTSGYEEENTTASG
DVGLAAIQIPKKGEDSKNNTKKEESDEEEEEEEEEENKESEAEVDKNKEVVNGT
DHGDGNGTSGKEDGEEEEEEVENGTESSVDKDNNSTELATSPNSRYESTTPEG
TYEVTSPPYEESTSNLYDSGYEEPVTNEYDNGYEIYENENGEPRGDNYRAYEDE
YSYKGYHPYNGYEGQDYYYHQ

Demonio de Tasmania

FSVKNLYRRAKLG DSEENGVLNHRHRYIYKYSYLYPPLKRIPVQRNGDSSEEGD
GESSEEEEEENSNEEEETESNGSEESTTLSATTLYGEENTTASGDIGLAAIQ
NPKKEEKNNQVTKKEESDEEEEEEEENESEAEVDENKEGGNSTNHENAEGDH
GNGNGGKEEGEEEEEVENGTEATENEASNSTEVTSLNGYEF TLLKGD FEVTS
PPYEESTSNLYDSGYEEPVTNEYDNGYEVYENENREPRGDNYRAYEDEYSYK
GHAYNGYGGQDYYYHQ

Anexo 3: Validación, sensibilidad y especificidad de servidores

Netphos

El servidor consta de dos predictores (sección 4.1.5): Netphos 2.0 y Netphosk 1.0 cuya validación fue realizada por los autores bajo el método de *cross-validation*. Dicho procedimiento consiste en dividir los grupos de datos en subgrupos de, aproximadamente, el mismo tamaño y usar uno de ellos como prueba mientras el resto son empleados para entrenar al sistema.

- Netphos 2.0: Se utilizaron secuencias, cuyos sitios de fosforilación hayan sido verificados experimentalmente, presentes en la *Phosphobase* (Kreepipuu, 1999). Los grupos de datos correspondientes a sitios de

fosforilación en Y, T y S, se dividieron en cinco subgrupos para realizar ser utilizados en la validación. Cuatro de los subgrupos se emplearon para entrenar a las NN y el restante se utilizó para probar las redes entrenada. Blom (1999) reporta lo siguiente:

- a) Predicción sobre serinas: sensibilidad, 50-90%, especificidad, 65-100%.
 - b) Predicción sobre treoninas: sensibilidad 65-87% y especificidad, 37-100%.
 - c) Predicción sobre tirosinas: sensibilidad 70-87% y especificidad, 40-100%.
- Netphosk 1.0: Se utilizaron secuencias, cuyos sitios de fosforilación hayan sido verificados experimentalmente, presentes en las bases de datos *Phosphobase* y *Swiss-Prot*. Los grupos de datos para cada enzima se dividieron en subgrupos para realizar la validación correspondiente. Uno de estos es utilizado para probar el desempeño de las NN, mientras que el resto de los subgrupos se usa para entrenar al sistema. Posteriormente estos subgrupos se rolan hasta que cada uno de ellos haya sido probado al menos una vez. Blom (2004) reporta lo siguiente:

Se realizó una comparación del método con las predicciones realizadas por el sitio *Scansite*. se encontró que a niveles dados de especificidad para ambos servidores (0.8,0.75,0.7,0.6 y 0.5 para NetphosK 1.0 y *high, medium* y *low* para *Scansite*) el método NetphosK 1.0 presenta una mayor sensibilidad en todos ellos. También se presentan los coeficientes de Mathews correspondientes, siendo de 0.32-0.58 para Netphos y de 0.13-0.26 para *Scansite*.

PONDR

Anteriormente se mencionó que PONDR integra tres predictores (sección 4.1.6): VL1, XC y XN. Los tres fueron sometidos al proceso de *cross-validation*.

Para VL1 se utilizaron cinco grupos de datos obtenidos de las bases de datos *SwissProt* y PIR. El 80% de ellos fue particionado de manera aleatoria: 64% fue utilizado para entrenar las NN y el 16% fue utilizado como grupo de prueba. Mientras

que el 20% restante se utilizó para estimar la generalización del predictor. No se reportan valores de sensibilidad o especificidad, pero se reporta la precisión del método con valores de $86\pm 2\%$ a $92\pm 1\%$.

Los predictores XC y XN fueron desarrollados con trescientas veintisiete proteínas diferentes no relacionadas entre sí, obtenidas de PDB-select-25. Se realizó el proceso de *cross-validation* por quintuplicado, sin embargo, no se da una explicación de cómo se dividieron los grupos de datos. Las precisiones reportadas son: $78.8\pm 1.5\%$ para regiones N terminales (predictor XN) y $75.3\pm 3.3\%$ para regiones C terminales (predictor XC).

PrDOS

El método utilizado por el servidor fue validado por Ishida (2007), mediante *cross-validation* quíntuple, no hay información respecto a los grupos de datos empleados. No hay información respecto a la sensibilidad o especificidad, pero en el 2006 la comunidad de biología estructural en la CASP7 (Evaluación Crítica de Estructura Secundaria, por sus siglas en inglés) evaluó el desempeño del predictor, donde demostró ser uno de los mejores métodos con una precisión $>90\%$.

Jpred

Para el algoritmo de predicción se utilizó un proceso de *cross-validation* séptuple, sin embargo, no se mencionan detalles sobre los grupos de datos empleados. La precisión reportada para predicciones es del 82%.

PSIPRED

La validación del método fue realizada por el proceso de *cross-validation* sin embargo no se dan más detalles al respecto. En sus inicios mostro una precisión del 76.5% y en el año de 1998 fue sometida a revisión en la CASP3 donde se posicionó como el mejor predictor con un valor de precisión de entre 73.4% y 77.3% en las diferentes categorías (McGuffin, 2000). Actualmente se han realizado mejoras al método dando como resultado una precisión de $81.4 \pm 0.6\%$ (Buchan et al., 2010).

CFSSP

Este método, en sus respectivas publicaciones, no da detalle alguno sobre algún proceso de validación, sensibilidad o especificidad; sin embargo, Kumar (2013) menciona que en la práctica real el método tiene una precisión de 50% - 60%.