



UNIVERSIDAD NACIONAL AUTÓNOMA DE MEXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS QUÍMICAS

Estudio estructural e interacción del dominio r1 del factor de transcripción A-Myb de humano

TESIS

PARA OPTAR POR EL GRADO DE

MAESTRO EN CIENCIAS

PRESENTA

Q. Salomón Pineda Silva

TUTOR

Dr. José Federico del Rio Portilla
Instituto de Química, UNAM.

CIUDAD DE MÉXICO, DICIEMBRE DE 2021



Universidad Nacional
Autónoma de México

Dirección General de Bibliotecas de la UNAM

Biblioteca Central



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.



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS QUÍMICAS

**Estudio estructural e interacción del dominio r1 del factor de transcripción A-Myb
de humano**

**TESIS
PARA OPTAR POR EL GRADO DE**

MAESTRO EN CIENCIAS

PRESENTA

Q. SALOMÓN PINEDA SILVA



Ciudad de México, diciembre de 2021.

Jurado asignado:

Presidente	Dra. Irma Bernal Lugo
Vocal	Dra. Elisa Irene Azuara Liceaga
Vocal	Dr. Alejandro Sosa Peinado
Vocal	Dra. Patricia Cano Sánchez
Secretario	Dra. Alejandra Hernández Santoyo

Sitio donde se desarrolló el proyecto:

Laboratorio 1, Departamento de Química de Biomacromoléculas, Instituto de Química,
Ciudad Universitaria

Agradecimientos

Agradezco inmensamente al Dr. Federico del Rio Portilla, por aceptarme en su equipo de trabajo, así como por brindarme su guía y apoyo durante esta etapa de mi formación académica y personal.

Agradezco a la Universidad Nacional Autónoma de México, a la Facultad de Química y al Instituto de Química por acogerme y darme la oportunidad de continuar desarrollándome académicamente.

Gracias también a mis compañeros de laboratorio, cuya experiencia profesional y asistencia fue de gran ayuda para la realización de este proyecto.

Agradezco sinceramente a cada uno de mis sinodales por el tiempo invertido durante la revisión de este trabajo: Dra. Irma Bernal Lugo, Dra. Elisa Irene Azuara Liceaga, Dr. Alejandro Sosa Peinado, Dra. Patricia Cano Sánchez, Dra. Alejandra Hernández Santoyo

Agradezco a mi familia, por estar conmigo en cada momento y alentarme a seguir adelante con mi formación académica.

Este proyecto fue posible gracias al apoyo otorgado por el Consejo Nacional de Ciencia y Tecnología (CONACyT) a través del programa de becas para estudios de posgrado, número de becario 997730.

Tabla de contenido

Resumen	8
1. Marco teórico.....	9
1.1. Proteínas.....	9
1.1.1. Estructura de proteínas	9
1.1.2. Proteínas recombinantes	11
1.2. Técnicas de caracterización de proteínas.....	11
1.2.1. Electroforesis en gel de poliacrilamida con SDS (SDS-PAGE).....	12
1.2.2. Espectrometría de masas MALDI-TOF	13
1.2.3. Dicroísmo circular	15
1.2.4. Resonancia magnética nuclear.....	16
1.3. Modelado Molecular y predicción de estructura.	18
1.3.1. Predicción por homología.....	19
1.3.2. Predicción <i>ab initio</i>	21
1.3.3. AlphaFold	22
1.4. Acoplamiento molecular	22
2. Antecedentes.....	24
2.1. Factores de transcripción y familia Myb	24
2.1.1. Dominio de unión con el ADN	27
2.1.2. Dominio R1	31
2.1.3. Secuencia de unión Myb-ADN:	32
3. Hipótesis	35
4. Objetivo general	35
Objetivos particulares	35
5. Metodología.....	36
5.1. Modelado Molecular.....	36
5.1.1. Obtención del modelo.....	36
5.1.2. Minimización de la energía y análisis y selección de modelos.	38
5.1.3. Análisis del modelo	40
5.2. Obtención de R1	44
5.2.1. Diseño del gen y transformación	44
5.2.2. Optimización de la expresión.	45
6. Resultados y análisis de resultados.....	47

6.1.	Modelado molecular	47
6.1.1.	Predicción de estructura secundaria y modelado molecular.....	47
6.1.2.	Minimización de la energía.	54
6.1.3.	Análisis del modelo	58
6.2.	Obtención de R1	73
6.2.1.	Transformación y expresión.	73
7.	Conclusiones y perspectivas	76
8.	Bibliografía y referencias	77

Resumen

Los factores de transcripción de las células eucariotas son clasificados en familias de acuerdo con los diferentes tipos de dominio de unión al ADN (DNA-Binding Domain, DBD). La diversificación de las funciones de estos factores comúnmente recae más en las otras partes de la proteína que sobre el DBD, dado que los miembros dentro de una familia frecuentemente reconocen la misma secuencia de ADN. Tal es el caso de la familia de factores de transcripción Myb, donde los tres miembros presentes en los vertebrados A-, B- y c-Myb comparten DBDs con secuencias de aminoácidos similares que se unen a la misma secuencia de ADN.

Los DBDs de las proteínas Myb de vertebrados se encuentran compuestos de tres repeticiones imperfectas, es decir, no idénticas, de 52-53 aminoácidos, denominadas repetidos R1, R2 y R3, cada uno con un motivo característico rico en triptófano. Los estudios sobre estos tres repetidos han determinado que la unión de R2 y R3 (designada R2R3) constituye un mínimo necesario para formar un DBD que se una a una secuencia específica de ADN, mientras que R1 parece no tener interacciones específicas con el ADN y su función no ha sido determinada.

En este proyecto se emplearon metodologías de mecánica molecular para determinar la estructura del fragmento R1 del dominio de unión al ADN de la proteína A-Myb de humano, así como sus interacciones intramoleculares. También se llevó a cabo un breve estudio de acoplamiento molecular para determinar la existencia de interacciones entre R1 con un sitio de unión de Myb.

Los resultados indicaron la presencia de un núcleo hidrofóbico que proporciona estabilidad al fragmento, los estudios de acoplamiento molecular con ADN no mostraron interacciones consistentes entre el ADN y el fragmento R1.

1. Marco teórico

1.1. Proteínas.

Las proteínas son biopolímeros macromoleculares formadas en cadena por unidades más pequeñas llamadas aminoácidos. Estos aminoácidos se encuentran unidos por enlaces de tipo amida, denominado también enlace peptídico (Figura 1), en el cual el grupo amino de una unidad forma un enlace mediante condensación con el carboxilo del siguiente aminoácido. A las unidades de aminoácidos en una cadena se les llama también residuos, mientras que a la secuencia repetida y extensa de átomos -N-CH-CO- que forman una cadena continua se le llama esqueleto de la proteína. Por convención, los péptidos se escriben con el aminoácido terminal N a la izquierda y el aminoácido terminal C a la derecha. Las proteínas cumplen una amplia variedad de funciones dentro de los organismos, incluyendo la catálisis de reacciones metabólicas, replicación del ADN, respuesta a estímulos, estructura celular y transporte molecular, entre otras. ¹.

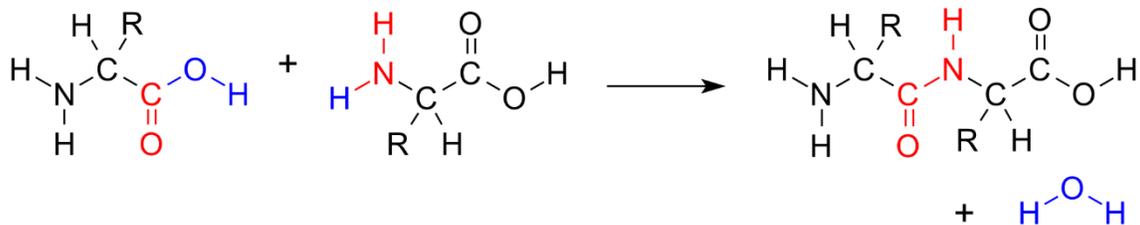


Figura 1. Formación de un enlace peptídico entre dos aminoácidos.

La estabilidad estructural de las proteínas determina si se encuentran en su estado plegado nativo o en su estado desnaturalizado.

1.1.1. Estructura de proteínas

La estructura de las proteínas está determinada por la secuencia de aminoácidos que la conforman y es estabilizada por varias interacciones entre átomos o grupos de átomos, estas interacciones pueden ser de tipo hidrofóbicas, electrostáticas, puente de hidrógeno, fuerzas

de *van der Waals* y enlaces disulfuro ². La estructura puede describirse en cuatro niveles diferentes, que son la estructura primaria, secundaria, terciaria y cuaternaria.

La **estructura primaria** es la estructura covalente que se conforma de la secuencia de aminoácidos más cualquier otra modificación covalente, incluyendo enlaces disulfuro, fosforilaciones, etc. Muchas de las propiedades de las proteínas derivan, directamente o indirectamente, de la estructura primaria ³.

La **estructura secundaria** es la conformación local de la proteína en ciertas partes de la cadena polipeptídica. Esta estructura es resultado de la interacción, mediante puentes de hidrógeno, de los átomos que forman el esqueleto de la proteína. Ciertos tipos de estructura secundaria son particularmente estables y se presentan ampliamente en un gran número de proteínas, siendo los elementos más comunes las hélices α , los motivos β , los giros β , las hélices de poli prolinas y hélices 3_{10} .

La **estructura terciaria** es el plegamiento tridimensional completo de la cadena polipeptídica. La estructura terciaria involucra todos los elementos de estructura secundaria de la cadena, así como los enrollamientos y pliegues que haya entre cada uno de ellos. A diferencia de la estructura secundaria, la estructura terciaria toma en cuenta las interacciones de las cadenas laterales, la estructura global es estabilizada por interacciones hidrofóbicas, interacciones con el medio externo, interacciones electrostáticas y puentes de hidrógeno.

Finalmente, la **estructura cuaternaria** es la forma en la que dos o más cadenas polipeptídicas plegadas se agrupan en complejos proteicos denominados oligómeros. La estructura cuaternaria involucra el número y la naturaleza de las distintas subunidades, o monómeros, y la forma en la que se asocian espacialmente. Las interacciones responsables de la formación de aglomerados son interacciones hidrofóbicas, hidrofílicas, e interacciones electrostáticas ³.

1.1.2. Proteínas recombinantes

Una proteína recombinante es aquella que es producida en un organismo que ha sido sujeto a modificación genética y que, por lo general, el organismo es diferente a aquel en el cual la proteína se produce de forma natural. La mayoría de las proteínas recombinantes tienen como finalidad el uso por humanos en distintos ámbitos, siendo los principales las aplicaciones en el área de la salud, así como en la industria alimenticia. Comúnmente, estas proteínas son expresadas en otros organismos, tales como bacterias, levaduras o cultivos de células animales; éstas usualmente son obtenidas a partir de genes aislados, o sintéticos, y clonados en un vector de expresión. El organismo más comúnmente empleado para la expresión de proteínas recombinantes es la *Escherichia coli*; esta bacteria ha demostrado ser un sistema de expresión robusto y eficiente en costo para la producción a gran escala de proteínas recombinantes. *E. coli* se encuentra bien caracterizada y es fácil de manejar y manipular genéticamente, su rápido crecimiento, medio de cultivo económico, altos niveles de expresión y la facilidad de escalar los procesos a niveles industriales proveen grandes ventajas para la producción de proteínas recombinantes con fines terapéuticos o estudios funcionales ⁴.

Para la expresión de genes en bacterias son empleados plásmidos especiales denominados vectores de expresión. Estos vectores proveen un promotor fuerte para dirigir la expresión del gen clonado; también contiene genes que proveen a la célula con resistencia a antibióticos para permitir la selección del vector y, por tanto, de células transformadas con el gen para producir la proteína recombinante. Además, deben tener un origen de replicación adecuado para la célula huésped ⁵.

1.2. Técnicas de caracterización de proteínas.

Tras y durante la obtención y purificación de una proteína a partir de su fuente natural o mediante métodos recombinantes, es necesaria la caracterización de la misma, la cual provee información sobre las características químicas y estructurales de la proteína. En los procesos de obtención, la caracterización permite optimizar la producción e incrementar el

rendimiento de los métodos de obtención, reduciendo el riesgo de la pérdida de recursos. Tras la obtención, es necesario verificar si la proteína ha sido obtenida correctamente, así como la presencia de moléculas adicionales y la estimación de la concentración de la proteína⁶. Entre las técnicas más comunes empleadas para la caracterización de proteína se encuentran la electroforesis en gel de poliacrilamida con SDS, la espectrometría de masas, el dicroísmo circular, la resonancia magnética nuclear y la cristalografía de rayos X. Recientemente se ha añadido también la técnica de criomicroscopía electrónica, especialmente en la determinación de estructuras de gran tamaño, tales como cápsides de virus o fibras amiloides. A continuación, se describen algunas de las técnicas mencionadas:

1.2.1. Electroforesis en gel de poliacrilamida con SDS (SDS-PAGE)

Los métodos de electroforesis permiten la separación de proteínas con base en la migración de proteínas cargadas en un campo eléctrico. Si bien estos métodos no permiten la purificación de proteínas en grandes cantidades, son especialmente útiles para visualizar proteínas. Con este método es posible estimar el número de proteínas diferentes en una mezcla o determinar el grado de pureza de una proteína. De igual forma, permiten determinar características cruciales de una proteína, tales como el punto isoeléctrico, así como estimar el peso molecular.

La electroforesis en gel de poliacrilamida con SDS (SDS-PAGE, Figura 2) consiste en la separación de proteínas por electroforesis en gel de poliacrilamida. Este gel consiste de una matriz porosa inerte, cuyo tamaño de poro depende del porcentaje de acrilamida durante su preparación. El gel actúa como una caladera molecular, alentando la migración de las proteínas en una proporción aproximada a su relación masa/carga. La migración de la proteína puede verse afectada por la forma de la misma. El movimiento de la macromolécula es impulsado por la presencia de un campo eléctrico a lo largo del gel. Para minimizar el efecto de la forma y la relación carga/masa se emplea el detergente dodecil sulfato sódico (SDS), el cual se une a la mayoría de las proteínas en cantidades aproximadamente proporcionales a su peso molecular. Esta unión contribuye una carga neta negativa a la proteína, generando que la carga interna de la misma sea insignificante en comparación, dándole a todas las proteínas una relación carga/masa similar. Adicionalmente, la

configuración nativa de una proteína es alterada con la unión de SDS y la mayoría de las proteínas asumen una forma similar. Debido a esto, la separación de proteínas en SDS-PAGE depende casi exclusivamente de su peso molecular, con los polipéptidos más pequeños migrando mucho más rápido.⁷

Tras la electroforesis, las proteínas son visualizadas mediante el uso de un tinte, el cual se une a la proteína, pero no al gel. La intensidad del color de las bandas es normalmente proporcional a la cantidad de proteína en el gel, esto permite estimar la pureza de la proteína y, empleando patrones de peso molecular conocido, la masa molecular de una proteína.⁶

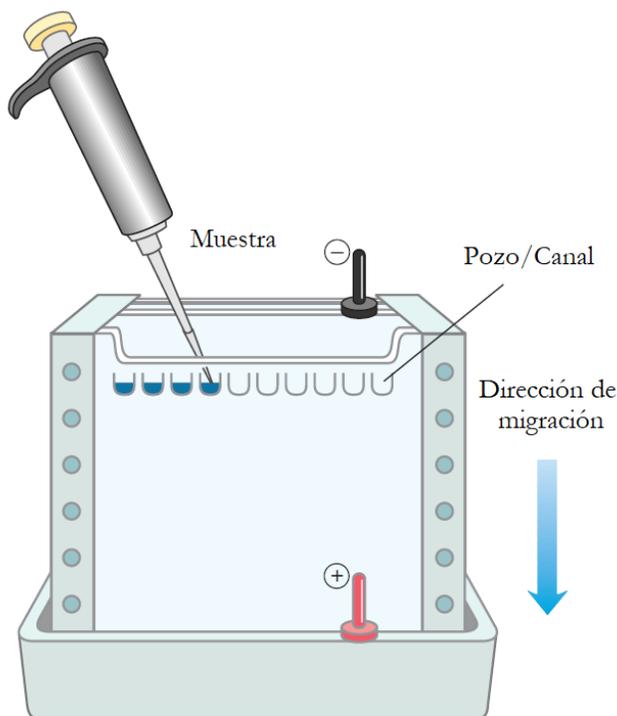


Figura 2. Diagrama de SDS-PAGE, diferentes muestras son cargadas en los pozos o canales en la parte superior del gel de poliacrilamida. Las proteínas entran al gel cuando un campo eléctrico es aplicado. El gel minimiza las corrientes de convección causadas por los gradientes de temperatura, al igual que los movimientos de las proteínas no inducidos por el campo eléctrico⁷.

1.2.2. Espectrometría de masas MALDI-TOF

MALDI-TOF (Matrix-Assisted Laser Desorption/Ionization Time-Of-Flight Mass Spectrometry – Desorción/Ionización mediante Láser asistida por Matriz) es una técnica

derivada de la desorción láser (LD) y permite llevar a cabo el análisis de espectrometría de masas de polímeros y macromoléculas de hasta cientos de miles de Daltons con un mínimo de fragmentación⁸.

El proceso MALDI permite obtener iones moleculares monocargados intactos, los cuales son difíciles de obtener mediante los métodos clásicos de ionización de espectrometría de masas (impacto electrónico e ionización química). El primer paso de este proceso consiste en colocar el analito en una matriz cristalina compuesta de especies orgánicas pequeñas de alta absorbancia, esta matriz es depositada en un soporte conductor previo al paso de desorción. Posteriormente, un pulso láser irradia la muestra generando la ablación y desorción de la misma y la formación de iones del analito.

La mezcla de analito con la matriz es preparada mediante la cocrystalización del analito con un exceso molar de las moléculas que forman la matriz, integrando las moléculas del analito a la red cristalina de forma homogénea. Esta mezcla permite obtener una forma eficiente y controlable para transferir energía a la muestra, mientras las moléculas de analito son protegidas del exceso de energía que normalmente llevaría a su descomposición. De igual forma, la dilución evita la asociación de las moléculas de analito, las cuales normalmente formarían complejos demasiado grandes para ser desorbidos y analizados⁸. La viabilidad de una molécula, o mezcla de moléculas, como matriz es dependiente de la muestra, y es, en muchos casos, determinada mediante ensayo y error. Si bien es posible determinar si una molécula candidata a matriz absorberá eficientemente energía de láser a una longitud de onda determinada, no es posible predecir si dicha molécula generará una solución sólida (o líquida) homogénea con las moléculas del analito, particularmente en los casos en los que se encuentran presentes contaminantes iónicos⁸.

Finalmente, los iones son acelerados a través de un campo eléctrico hasta alcanzar al detector. El espectrómetro de masas TOF determina la masa de los iones midiendo su velocidad para determinar el valor de la relación masa-carga (m/z). La velocidad de los iones es proporcional a $(m_i/z_i)^{-1/2}$, donde m_i/z_i es la relación masa carga de una especie iónica en particular. Durante el recorrido a través del campo eléctrico, los iones son separados en grupos

individuales de iones, cada uno viajando con una velocidad característica de su masa. Una señal es producida cada vez que un grupo de iones golpea el detector. La diferencia entre el tiempo de ionización, común para todos los iones, y la detección de un ion individual es proporcional a m/z y puede ser usada para determinar la masa iónica^{8,9}.

1.2.3. Dicroísmo circular

El dicroísmo circular (DC) es una técnica que permite determinar si una proteína expresada y purificada se encuentra plegada. También permite obtener información sobre su estructura secundaria y, bajo condiciones especiales, también sobre la estructura terciaria. Esta técnica hace uso de la diferente absorción de la luz polarizada circularmente, a la derecha o a la izquierda, por parte de compuestos o estructuras ópticamente activos, tales como las moléculas quirales o elementos de estructura secundaria. Las partes de las moléculas o de las estructuras responsables de absorción de luz polarizada son denominados cromóforos. En el caso de las proteínas plegadas, los cromóforos que interactúan con la luz polarizada son los enlaces peptídicos, cadenas laterales aromáticas, y enlaces disulfuro. Si algún número de cromóforos del mismo tipo se encuentran cerca, estos pueden comportarse como una sola unidad absorbente de luz polarizada, esto ocurre con los enlaces peptídicos en los elementos de estructura secundaria.¹⁰

En los equipos de dicroísmo circular la luz pasa a través de un polarizador para generar radiación polarizada, la cual posteriormente es convertida en luz circularmente polarizada. Los equipos generan luz polarizada en ambas direcciones alternando rápidamente. Cuando esta luz pasa a través de las especies ópticamente activas, éstas absorben la luz en diferente grado dependiendo de la dirección de la misma (hacia la derecha o hacia la izquierda), esta diferencia de absorbancia puede ser detectada y cuantificada³. Para la cuantificación de estructuras secundarias se emplea dicroísmo circular con radiación UV, la cual es absorbida por proteínas en longitudes de ondas cercanas a 200 nm, dicha longitud no permite detectar la quiralidad de los aminoácidos individuales, pero interactúa con los elementos de estructura secundaria. Un espectro de DC (Figura 3) puede ser descompuesto para obtener las fracciones relativas de estructuras de alfa-hélice, beta-plegada y vueltas. La cantidad de fracciones

relativas depende de la forma de los espectros de referencia para una proteína con 100% de esas estructuras, por tanto, es difícil obtener fracciones exactas, pero los valores aproximados son igualmente útiles.

DC es un proceso simple y ampliamente utilizado para verificar si una proteína se encuentra plegada, detectar cambios estructurales, y para medir la cinética de plegamiento. Cambios en el espectro alrededor de 200 nm indican cambios en la estructura secundaria, mientras que cambios en el área cercana a 280 nm indican cambios en el ambiente de residuos de Trp y Tyr y, por tanto, cambios en la estructura terciaria³.

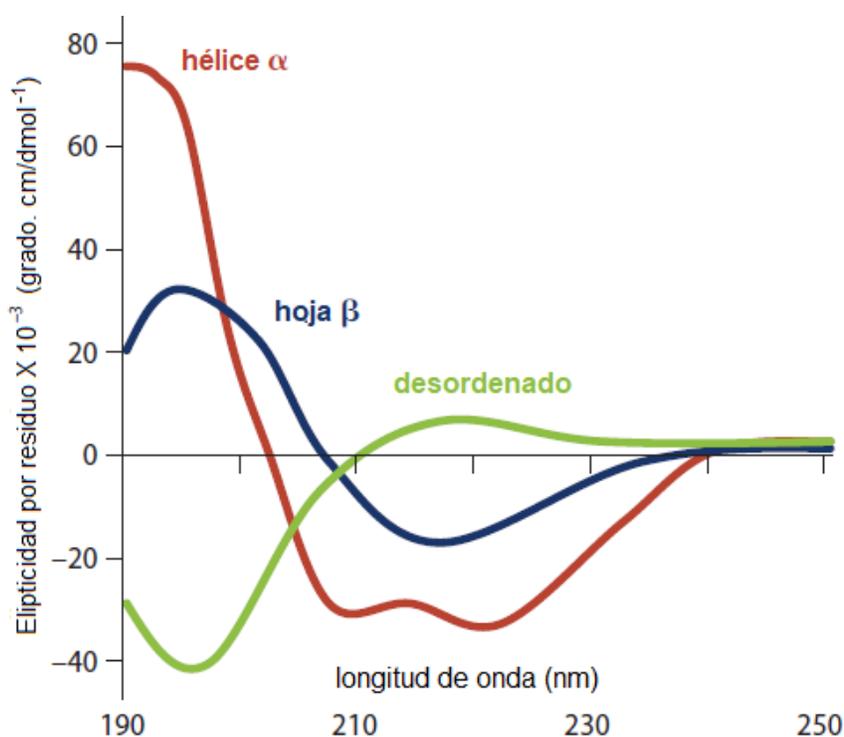


Figura 3. Espectro de dicroísmo circular típico para las principales estructuras secundarias de proteínas (imagen tomada de Williamson, M. 2012)³

1.2.4. Resonancia magnética nuclear.

La Espectroscopia de Resonancia Magnética Nuclear es capaz de revelar la estructura atómica de macromoléculas en disolución siempre y cuando se puedan obtener disoluciones lo suficientemente concentradas (~1 mM, o 15 mg mL⁻¹ para una proteína de 15 kD). Esta

técnica se basa en el espín nuclear. Sólo un limitado número de isótopos presentan magnetismo en el núcleo atómico, esta propiedad es llamada momento angular de espín. El más simple de estos núcleos es el núcleo de hidrógeno (^1H), el cual es un protón. El giro de los protones genera un momento magnético, inicialmente sin dirección específica. Cuando un campo magnético externo se aplica, el momento magnético del protón puede alinearse en dos posibles orientaciones o estados de espín: a favor del campo magnético (alfa) o en contra del campo magnético (beta). La diferencia energética entre esos estados es proporcional al campo magnético aplicado.

El estado alfa del momento magnético tiene una energía ligeramente menor a la del estado beta y por lo tanto su población es mayor. Un núcleo en estado alfa puede excitarse al estado beta aplicando un pulso de radiación electromagnética en forma de radiofrecuencia, siempre y cuando la frecuencia de la radiación corresponda a la diferencia energética entre los estados alfa y beta, en dichas circunstancias el espín cambiará entre los estados alfa y beta dando como resultado la resonancia. Un espectro de resonancia para una molécula puede obtenerse variando el campo magnético a una frecuencia constante de radiofrecuencia o variando la frecuencia de la radiación electromagnética a un campo magnético constante ¹¹. Estas propiedades pueden emplearse para examinar el ambiente químico de los núcleos en resonancia, el flujo de electrones alrededor de los núcleos magnéticos genera un campo magnético local que se opone al campo magnético aplicado. El grado de apantallamiento depende de la densidad electrónica alrededor del núcleo, por tanto, núcleos en diferentes ambientes entrarán en resonancia a diferentes intensidades de campo magnético o radiofrecuencias, dependiendo del tipo de experimento.

El núcleo de la muestra perturbada absorberá radiación electromagnética a una frecuencia que pueda ser medida; las diferentes frecuencias, denominadas desplazamiento químico, son expresadas en unidades fraccionarias delta (partes por millón, o ppm) relativas al desplazamiento de un compuesto estándar, normalmente tetrametilsilano para ^1H y ^{13}C , añadido a la muestra. En el trabajo con proteínas, la información obtenida a través de espectros de RMN de ^1H hace posible deducir cambios a grupos químicos específicos bajo diferentes condiciones, tales como el cambio conformacional de una proteína de un estado

desordenado a un elemento de estructura secundaria en respuesta a un cambio de pH. Es posible obtener más información examinando cómo el espín de un núcleo afecta a los núcleos vecinos. Induciendo magnetización en una muestra a través de la aplicación de un pulso de radiofrecuencia, es posible alterar el espín de un núcleo y examinar los efectos en el espín de los núcleos vecinos. Esta técnica es extremadamente útil para la elucidación de la estructura de proteínas ya que permite obtener espectros de dos dimensiones que muestran la relación entre los núcleos ¹¹.

La Espectroscopia de Correlación Total, TOCSY (Total Correlated Spectroscopy), permite correlacionar protones que se encuentren en un mismo sistema de espín, estos son protones que se encuentran acoplados directa o indirectamente entre ellos. En el caso de una proteína, cada aminoácido pertenece a un sistema de espín diferente y TOCSY permite determinar la presencia de los tipos de aminoácidos que corresponden al sistema de espín detectado. El experimento de Espectroscopia de Correlación COSY (Correlation Spectroscopy) permite observar las correlaciones ¹H-¹H mediante acoplamiento escalar entre protones a no más de tres enlaces de distancia. Por último, la Espectroscopia de Efecto Nuclear Overhauser NOESY (Nuclear Overhauser Enhancement Spectroscopy) permite detectar la posición relativa de átomos, usualmente protones, a una distancia máxima de 5 Å. Esto permite reconstruir la estructura plegada de la proteína a partir de las interacciones entre protones cercanos.

1.3. Modelado Molecular y predicción de estructura.

El modelado molecular consiste en el desarrollo y uso de metodologías teóricas y experimentales para modelar, imitar y predecir el comportamiento de moléculas, iones o partículas, desde de sistemas químicos pequeños, como iones y átomos, hasta sistemas macromoleculares, tales como macromoléculas biológicas y conjuntos de materiales.¹²

En el estudio de las proteínas una de las áreas donde más se involucra el modelado molecular es la predicción de estructuras tridimensionales a partir de la secuencia de aminoácidos. Descifrar la conformación nativa de una proteína a partir de su secuencia es uno de los

desafíos más antiguos en biología molecular y computacional. Varios métodos han sido propuestos para alcanzar dicho objetivo, los cuales han sido evaluados por el foro CASP: Evaluación crítica de las técnicas para la predicción estructural proteica (*Critical Assessment of Techniques for Protein Structure Prediction*) usando predicciones de doble ciego. Los métodos para la predicción de estructura clásicos se pueden dividir en dos categorías principales de acuerdo con el enfoque utilizado: predicción por homología y predicción *ab initio*.¹³ Recientemente se ha desarrollado un nuevo método de predicción de estructuras llamado *AlphaFold*. Este método emplea inteligencia artificial y ha demostrado un gran potencial para la predicción de estructuras de proteínas, principalmente en los casos donde no se cuenta con estructuras homologas resueltas.¹⁴

1.3.1. Predicción por homología

El modelado por homología es la construcción de un modelo tridimensional de una proteína a partir de su secuencia y de la estructura obtenida experimentalmente de una proteína homóloga, la cual es denominada “molde” o “plantilla”. Se basa en el principio de que los elementos de estructura se encuentran más conservados que las secuencias. Un modelo molecular de una proteína se puede construir asignando una conformación con base en un alineamiento de secuencias, seguido de la construcción del modelo y la minimización de la energía.¹³

Dos conceptos importantes en el modelado por homología son la identidad y la similitud entre dos secuencias proteicas. La **identidad** de una secuencia con respecto a otra es la cantidad de residuos que coinciden exactamente entre ambas. Los espacios no son tomados en cuenta y el resultado es reportado con relación a la secuencia de menor tamaño, esta propiedad no es transitiva, si la secuencias A=B y B=C, esto no significa que A=C. La **similitud** es una descripción general de la relación entre dos secuencias. Considera la sustitución de aminoácidos con propiedades fisicoquímicas similares entre secuencias (sustitución conservativa) La medición de la similitud tienen como criterio la relación entre los aminoácidos no idénticos. Un reporte de similitud incluye la identidad de las secuencias y los aminoácidos similares.

El método general para la obtención por homología de la estructura tridimensional de una proteína es:

- i. Para una secuencia objetivo, identificar un molde adecuado empleando BLAST (Basic Local Alignment Search Tool - Herramienta de búsqueda de alineaciones locales básicas).
- ii. Alineamiento de secuencias.
- iii. Corrección del alineamiento para asegurarse que los residuos conservados o funcionalmente importantes se encuentren alineados.
- iv. Generar el esqueleto del modelo.
- v. Modelado en bucles.
- vi. Modelado de las cadenas laterales empleando bibliotecas de rotámeros.
- vii. Optimización del modelo mediante minimización de energía.
- viii. Validación del modelo mediante la evaluación estereoquímica empleando el gráfico de Ramachandran.

La proteína molde debe tener la mayor identidad posible con la proteína objetivo. Es igualmente importante considerar el ambiente y la calidad de la estructura de la proteína molde. Generalmente se busca una identidad mayor al 40%. Identidades menores requieren pasos adicionales para su modelado.

Las principales limitantes de este método son:

- El nivel de conservación estructural entre la proteína plantilla y la consulta son puntos determinantes de la calidad del modelo resultante. El modelo requiere la existencia de una proteína con alta similitud y estructura conocida. La calidad de esta estructura es igualmente importante.
- El software actual requiere un incremento de capacidad.
- La conformación de las cadenas laterales no es fácil de determinar.

1.3.2. Predicción *ab initio*.

La predicción mediante métodos *ab initio* tiene como objetivo predecir la estructura de la proteína basándose únicamente en la secuencia de esta y sin ningún molde o plantilla. Los métodos *ab initio* consideran todas las interacciones de enlace y no enlace involucradas en el proceso de plegamiento y eligen la estructura con la menor energía libre. Este enfoque se basa en la “hipótesis termodinámica”, la cual establece que la estructura nativa de una proteína es aquella en la cual la energía libre se encuentra en el mínimo global o cerca del mismo.¹³

Los métodos *ab initio* son normalmente usados cuando estructuras conocidas de secuencias homólogas (e incluso homólogas pobres) no se encuentran disponibles. Estos métodos dividen el problema de encontrar conformaciones nativas o cercanas a las nativas a partir de una secuencia en dos problemas menores: (i) crear una función que pueda distinguir entre la estructuras correctas e incorrectas y (ii) desarrollar un método de búsqueda para explorar el espacio conformacional

El protocolo básico de los métodos *ab initio* para la predicción de estructura inicia suministrando la secuencia de aminoácidos que conforman la proteína. Posteriormente se inicia la búsqueda de diferentes conformaciones posibles hasta la obtención de una estructura viable como estructura plegada nativa, esta búsqueda comúnmente emplea métodos Monte Carlo para abarcar más espacios conformacionales, reduciendo el tiempo de cómputo. Sin embargo, existen dos problemas que han limitado la eficacia de dicho enfoque. Primero, los métodos de dinámica molecular son muy exigentes en términos de requerimientos computacionales debido al amplio espacio conformacional a cubrir y a la complejidad de las funciones de energía. El segundo problema es la falta de funciones de potencial adecuados para sistemas de macromoléculas en agua, este problema es considerado el más serio y limitante de los dos.¹⁵

1.3.3. AlphaFold

AlphaFold es un método computacional capaz de predecir la estructura de proteínas con gran precisión, incluso en casos donde no existan estructuras similares conocidas. Este método fue puesto a prueba en la Catorceava Evaluación crítica de las técnicas para la predicción estructural proteica (CASP14), donde las estructuras generadas fueron más precisas que las generadas por los otros competidores. AlphaFold mejora la precisión de la predicción de estructuras mediante la incorporación de arquitecturas de redes neurales y un proceso de “entrenamiento” basado en las restricciones evolutivas, físicas y geométricas de las estructuras de proteínas.¹⁴

1.4. Acoplamiento molecular

El acoplamiento molecular es el estudio de como dos estructuras moleculares se acoplan juntas, comúnmente llamado Docking en inglés. Se trata de una técnica de análisis molecular empleada para predecir como una molécula grande (normalmente proteínas o ácidos nucleicos) interactúa con una molécula de menor tamaño. La habilidad de proteínas, tales como enzimas, de interactuar con moléculas pequeñas para formar complejos supramoleculares juega un rol importante en la dinámica de proteínas, el cual puede inhibir o aumentar su función biológica. El comportamiento de las moléculas pequeñas (denominadas ligandos) en los sitios de unión puede ser descrito mediante acoplamiento molecular. El método tiene como objetivo identificar la orientación espacial del ligando en el sitio de interacción con la proteína y predecir la afinidad entre los ligandos y la macromolécula. De acuerdo con el tipo de ligando, el acoplamiento puede ser clasificado como:

- Acoplamiento proteína-molécula pequeña (ligando)
- Acoplamiento proteína-ácido nucleico.
- Acoplamiento proteína-proteína.

De estas, el acoplamiento proteína-molécula pequeña es la que presenta la menor complejidad. Por otro lado, los acoplamientos proteína-proteína son típicamente los más

complejos, esto debido a la alta flexibilidad de las proteínas y su vasto espacio conformacional.¹⁶

El resultado del proceso de acoplamiento molecular depende del algoritmo de búsqueda y consiste en dos pasos. El primero es la generación de todas las posibles conformaciones y orientaciones de la molécula objetivo con el ligando. El segundo paso consiste en la evaluación de las interacciones favorables de cada una de estas combinaciones. Tras obtener las afinidades, todos los modelos son ordenados de acuerdo con el puntaje obtenido en el paso de evaluación.¹⁶

2. Antecedentes.

2.1. Factores de transcripción y familia Myb

Los factores de transcripción (FT) son proteínas que se unen a una secuencia particular del ADN en regiones reguladoras y que controlan la transcripción de los genes. La regulación de la expresión genética a nivel de la transcripción es un mecanismo de control ubicuo en muchos procesos biológicos, como el crecimiento, la diferenciación celular y el metabolismo. Los FT pueden actuar como activadores o supresores de la transcripción de su(s) gen(es) blanco(s). Estos factores son clasificados en familias de acuerdo a la estructura que emplean para mediar su unión con el ADN o para causar la dimerización de factores, la cual es frecuentemente un paso esencial en la unión con el ADN, algunos ejemplos son los factores con dedos de zinc, hélice-giro-hélice, cremallera de leucina y hélice-bucle-hélice, entre otros^{17,18}. Dentro de la familia de FT hélice-giro-hélice se encuentran los factores de transcripción tipo Myb. El primer gen *myb* que se identificó fue el oncogén *v-myb* del virus de mieloblastosis aviar, más tarde se descubrió que las proteínas transformantes *v-Myb* codificadas por estos virus son versiones truncadas, mutadas y expresadas constitutivamente de C-Myb, proteína producto del protooncogén *c-Myb*¹⁹; desde entonces se han identificado diversas proteínas con dominios Myb en los protozoarios, hongos, algas, plantas y animales.¹⁷

La familia de proteínas Myb pueden ser divididas en dos tipos básicos, el primero lo constituyen *c-Myb* (y sus derivadas *v-Myb*), *A-Myb* y *B-Myb*, estas últimas son dos proteínas que están altamente relacionadas con *c-Myb* y tienen presencia en organismos vertebrados (Figura 4). En general estas tres proteínas tienen dominios de unión con el ADN con secuencias de aminoácidos y con secuencias de reconocimiento similares. El otro tipo de proteínas Myb lo conforman proteínas con una mayor divergencia en sus dominios de unión con el ADN comparadas con las del primer grupo y han sido principalmente identificadas en hongos y plantas. Todas las proteínas Myb se localizan en el núcleo y se unen con el ADN por lo cual se presume que actúan regulando la expresión de genes involucrados en el control

de crecimiento y diferenciación, mediando la interacción entre secuencias específicas de ADN y otros componentes proteicos de la maquinaria de transcripción^{19,20}.

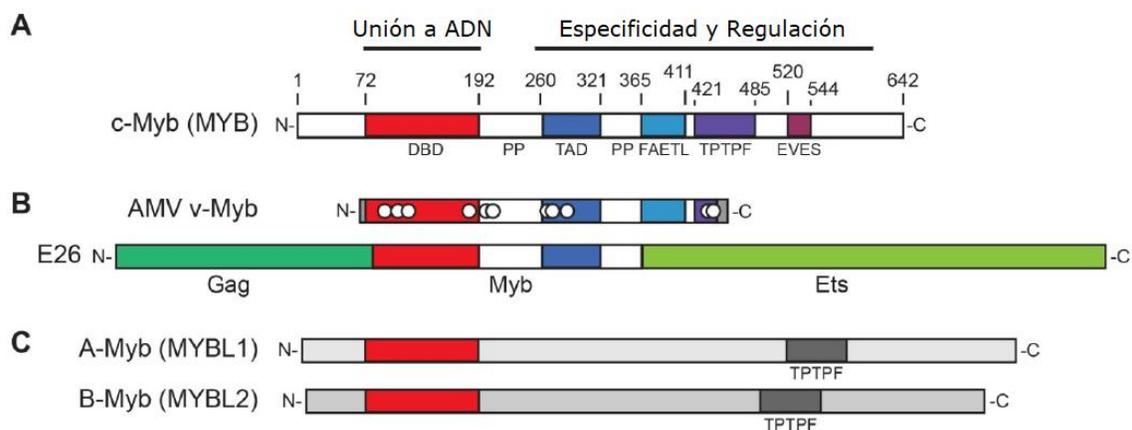


Figura 4. Estructuras de las proteínas Myb y sus dominios conservados. A) Dominios conservados en c-Myb. En el diagrama de c-Myb, las regiones conservadas en humanos, ratones y pollos se encuentran coloreadas, y la posición de los residuos se indica en la parte superior, el dominio de unión al ADN se indica en color rojo cerca de la sección N-terminal. Se indican, además, otros dominios conservados, tales como el dominio de activación transcripcional (TAD), requerido para la activación de la expresión de genes, el dominio FAETLI, requerido para la actividad oncogénica, el dominio TPTPFI, que se conserva también en otras proteínas Myb, el dominio EVES, responsable de interacciones intramoleculares y regulación negativa, y PP, una zona rica en prolina. B) Estructuras del gen oncogénico v-Myb producidos por el virus de la mieloblastosis aviar (AMV) y el virus E26. Los puntos blancos en el diagrama de v-Myb de AMV indican mutaciones que la diferencian de c-Myb. C) Diagramas de las estructuras de A-Myb y B-Myb. Estas proteínas son muy similares a c-Myb en su dominio de unión al ADN y en el dominio TPTPF.²¹

La estructura general de las proteínas presentes en vertebrados (A-Myb, B-Myb y c-Myb) consta de tres partes: Primero, un dominio de unión con el ADN en la región N-terminal, este dominio de unión consta de tres secuencias repetidas características (R1, R2 y R3); en la parte central de la proteína se ha identificado un dominio de activación transcripcional (TAD), el cual presenta una naturaleza hidrofílica; finalmente, en la región C-terminal, se encuentra un dominio regulador negativo, el cual tiene como función disminuir la actividad del dominio TAD.^{17,22}

La similitud de las proteínas Myb de vertebrados no termina únicamente con la homología estructural, sino que se extiende a su funcionalidad: la expresión de las proteínas A-, B- y c-Myb se asocia con la proliferación de tejidos. En general, los FT Myb se han implicado en la regulación de procesos fundamentales tales como el control de la proliferación y diferenciación celular, la regulación del metabolismo y la señalización celular en respuesta a

estímulos externos. Además, estudios han determinado que cuando las células terminan su proceso de diferenciación y entran a un nuevo ciclo celular, los genes Myb se desactivan.

La diferencia entre cada una de ellas radica en sus respectivos patrones de expresión, los cuales varían para cada proteína. Son estos mecanismos de expresión específicos para cada variedad los responsables de la diversidad de las funciones de las tres proteínas Myb en los vertebrados antes y después del nacimiento del organismo.²³

La primera proteína miembro de esta familia, c-Myb se expresa principalmente en células hematopoyéticas inmaduras, donde juega un papel fundamental para mantener el balance entre la proliferación y la diferenciación, así como en otros linajes celulares inmaduros de tejidos de rápida proliferación tales como los folículos capilares y el epitelio gastrointestinal. La sobreexpresión de c-Myb inhibe la diferenciación del precursor de células hematopoyéticas, mientras que la pérdida de expresión inhibe la proliferación y causa falla de la hematopoyesis intrafetal lo que lleva a la muerte. Estudios recientes han clasificado a c-Myb como un factor de transcripción pionero, pues se ha demostrado que puede asociarse con cromatina cerrada independientemente de otros factores y, por tanto, es capaz de modular la accesibilidad a la cromatina, lo cual ha incrementado el interés en su estudio²⁴. A-Myb es expresada principalmente en el sistema nervioso central durante su desarrollo, también se encuentra presente en la línea germinal de los linfocitos B, en el epitelio de los conductos de las glándulas mamarias y en los testículos. A-Myb regula de manera positiva la proliferación celular, en experimentos hechos con ratones se demostró que la inactivación de A-Myb causa un bloqueo en el desarrollo de la glándula mamaria y una espermatogénesis restringida. B-Myb es expresada de forma ubicua y regula de forma positiva el crecimiento y la diferenciación celular; la inactivación de B-Myb en ratones produjo que éstos murieran en etapas tempranas de desarrollo^{17,20}. La especificidad de la expresión de las proteínas Myb en diferentes tejidos indica la posibilidad de que se encuentren involucradas en diferentes rutas de regulación²⁵.

En la última década, la implementación de tecnologías avanzadas de biología molecular *in vitro* e *in vivo* han revelado la participación del oncogén *myb* en la carcinogénesis de

diferentes tipos de cáncer. Las proteínas Myb juegan un papel importante en la iniciación y mantenimiento de diferentes tipos de cáncer tales como la leucemia, cáncer de colon, cáncer de mama y el carcinoma adenoide quístico. En general, los niveles altos de expresión de Myb se asocian con el bloqueo en la diferenciación celular y una proliferación continua, lo cual deriva en oncogenicidad. La sobreexpresión de Myb se puede observar en todos los casos de leucemia mieloide aguda y linfocítica aguda, generalmente se asocia con un mal pronóstico de la enfermedad. En la mayoría de los casos, esta sobreexpresión es resultado de la sobre regulación de la transcripción del gen *myb*, lo cual a su vez se debe a una abundancia de oncoproteínas o a la translocación cromosomal, donde *myb* es colocado en la vecindad de otras secuencias regulatorias hiperactivas ²⁵.

Debido a esto, en tiempos recientes, se ha planteado la posibilidad de usar a los factores de transcripción Myb como blancos para los fármacos en el tratamiento del cáncer, lo cual ha abierto nuevos horizontes para la investigación. Estudios han identificado y caracterizado múltiples proteínas cofactores en el proceso de regulación de expresión genética, como posibles blancos en el tratamiento de cáncer, la inhibición de estas proteínas ha mostrado resultados prometedores tanto en estudios *in vitro* como *in vivo*, al grado de que se ha podido llegar a la fase de pruebas clínicas en humanos ²⁵. Myb es un blanco adecuado para la terapia tumoral debido a que su expresión anormal es clave para el crecimiento y desarrollo de múltiples tumores sólidos y leucemias. Un mayor entendimiento de la regulación de la expresión y actividad de los FT Myb ha derivado en el desarrollo de nuevas terapias con inhibidores de bajo peso molecular. Estudios *in vitro* han demostrado que múltiples inhibidores son efectivos para detener el crecimiento de tumores con elevada expresión de Myb y, dado que Myb juega un papel importante en la tumorigénesis, nuevos y más efectivos inhibidores contribuirán a incrementar la supervivencia de pacientes con tumores relacionados a su expresión ²⁶.

2.1.1. Dominio de unión con el ADN

Como ya se ha mencionado, la característica principal de todas las proteínas Myb es su dominio de unión con el ADN, ubicado típicamente cerca de la región N-terminal. Este

atributo es común entre las proteínas Myb de animales, plantas, protozoarios y hongos. (Figura 5) Un dominio Myb mínimo consiste en dos regiones de aproximadamente 50 aminoácidos de longitud, este dominio se encarga de mediar la unión de la proteína Myb al ADN, y se encuentra altamente conservado evolutivamente. Gran parte de las proteínas Myb presentan tres secuencias repetidas imperfectas (designadas como R1, R2 y R3) en estas regiones, aunque también se han identificado proteínas con cuatro o más segmentos repetidos. En el caso de las Myb presentes en vertebrados, la proteína prototipo c-Myb y las proteínas relacionadas A-Myb y B-Myb solo contienen las secuencias repetidas R1, R2 y R3.^{17,22} En contraste, las proteínas v-Myb y las proteínas Myb presentes en plantas y otros eucariontes han retenido únicamente las últimas dos secuencias (R2 y R3)¹⁹. Más aún, análisis de mutantes por eliminación de fragmentos de c-Myb determinaron que R2 y R3 son suficientes para el reconocimiento de las secuencias de ADN específicas y R1 solo incrementa la estabilidad del complejo Myb-ADN²⁷.

Cada una de las secuencias repetidas de Myb se encuentra compuesta de tres hélices alfa conectadas, en cada caso la tercera hélice se encuentra en contacto con el surco mayor del ADN haciendo que contactos sitio-específicos sean importantes para el reconocimiento de secuencias. Algo importante en las secuencias repetidas R2 y R3 es que se encuentran extremadamente cerca, lo que resulta en un traslapamiento parcial de las dos hélices de reconocimiento, alineándolas una tras otra en la hendidura mayor de la doble hélice; a pesar de ello, R2 y R3 cubren sólo una cara de la cadena de ADN, dejando amplio espacio para la interacción de la cadena de ADN con otras proteínas (Figura 6).¹⁹ Otros estudios comparativos entre la estructura de R1R2R3 libre y el complejo R1R2R3-ADN mostraron que R1 se une pobremente al ADN, sugiriendo que R1 se une y separa rápidamente del ADN²⁷.

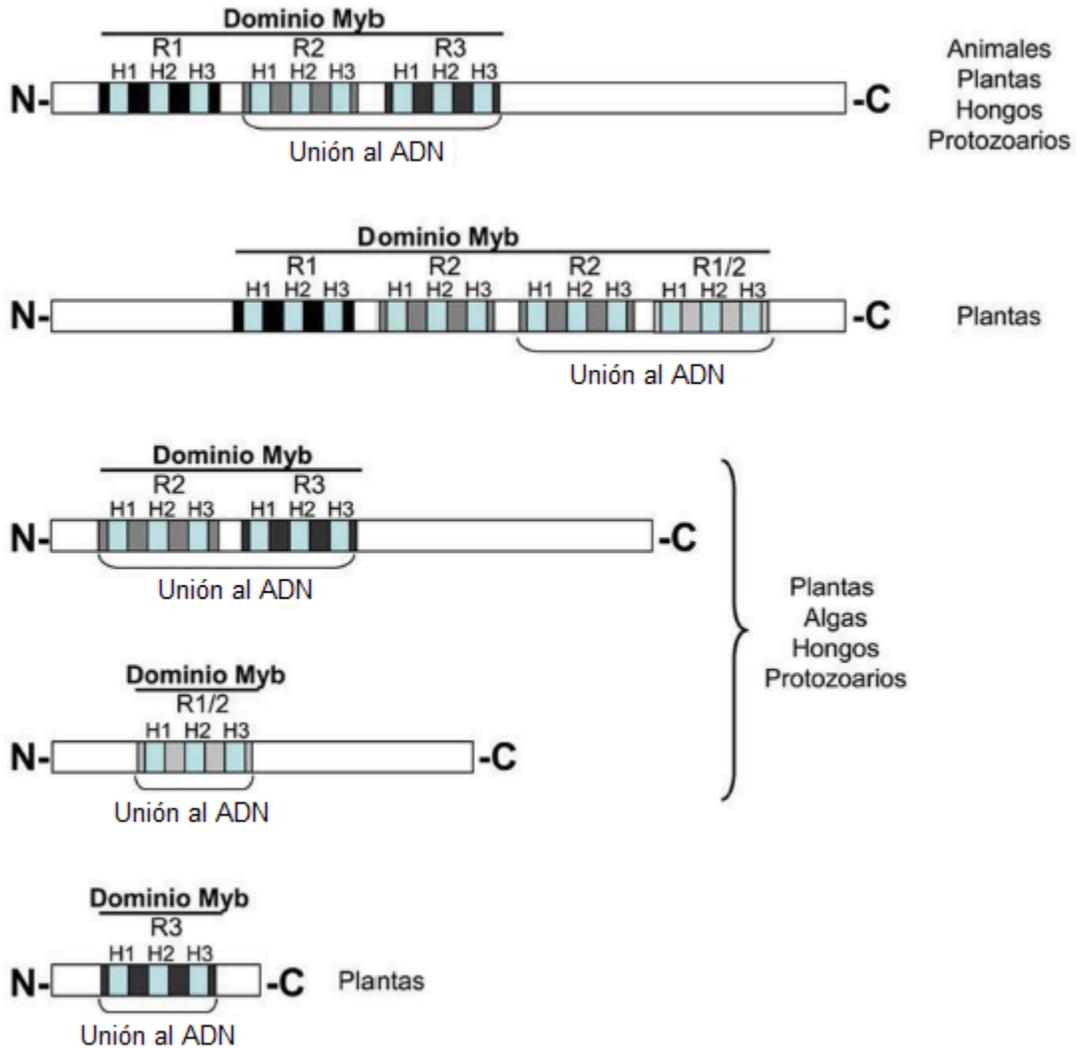


Figura 5. Clases de proteínas Myb. El esquema muestra las diferentes clases de proteínas Myb presentes en los organismos eucariontes dependiendo del número de secuencias repetidas, señaladas por R1, R2, R3. Se señalan las hélices de la estructura secundaria (H1, H2, H3) y el dominio de unión al ADN¹⁷.

Un factor clave en estas secuencias repetidas es la presencia periódica de triptófanos. Cada una de las tres series tiene tres triptófanos, los cuales se encuentran separados por 18 o 19 aminoácidos. Esta característica se puede observar en la c-Myb de ratones, humanos, pollos y *Drosophila*, así como en las secuencias de A-Myb y B-Myb del maíz y levadura (Tabla 1).

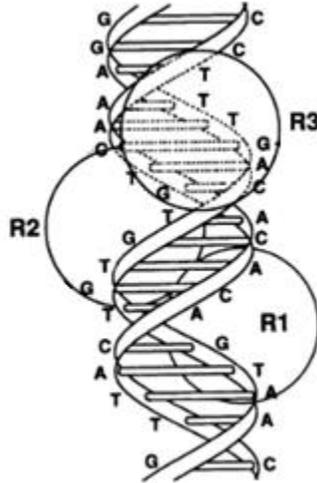


Figura 6. Modelo esquemático del complejo Myb-ADN. Se muestra el tamaño proporcional de cada uno de los fragmentos repetidos de Myb. Adaptado de Tanikawa et al., 1993

Tabla 1: Comparación de las secuencias del dominio de unión Myb de varias especies

	R1
Humano	LGKTRW T TREEDEKLK K LVEQNGTDD W KVIANYLPNRTDVQCQHR W QKVLNPE
Ratón	LGKTRW T TREEDEKLK K LVEQNGTDD W KVIANYLPNRTDVQCQHR W QKVLNPE
Gallo	LGKTRW T TREEDEKLK K LVEQNGT E D W KVIASFLPNRTDVQCQHR W QKVLNPE
Mosca	GFGKR W SKSE D VLLKQLVETHG-EN W EIIGPHFKDRLEQQVQ R WAKVLNPE
	R2
Humano	LIKGP W TKEEDQRVIELVQKYGPKR W SVIAKHLKGRIGKQCRER W HNHLNPE
Ratón	LIKGP W TKEEDQRVIELVQKYGPKR W SVIAKHLKGRIGKQCRER W HNHLNPE
Gallo	LIKGP W TKEEDQRVIELVQKYGPKR W SVIAKHLKGRIGKQCRER W HNHLNPE
Mosca	LIKGP W TRDEDDMVIKLVRFNFGPK W TLIARYLNGRIGKQCRER W HNHLNPN
	R3
Humano	VK K TS W T E E E DRIIYQ A HKRLGN R W A EIAKLLPGRTD N AIK N HW N ST M RR K V
Ratón	VK K TS W T E E E DRIIYQ A HKRLGN R W A EIAKLLPGRTD N AIK N HW N ST M RR K V
Gallo	VK K TS W T E E E DRIIYQ A HKRLGN R W A EIAKLLPGRTD N AIK N HW N ST M RR K V
Mosca	IK K TA W T E KEDEI I YQ A HLELGN Q W A KIAKRLPGRTD N AIK N HW N ST M RR K Y

Estudios de mutagénesis han sugerido que los residuos de triptófano no interactúan directamente con las bases del ADN, sino que forman un andamio o núcleo hidrofóbico que mantiene el plegamiento del motivo Hélice-giro-Hélice de los fragmentos R2 y R3, permitiendo que las regiones adyacentes, ricas en residuos básicos, interactúen con el ADN

2.1.2. Dominio R1

El análisis por eliminación de fragmentos han demostrado que únicamente R2 y R3 son absolutamente necesarias para la formación del complejo Myb con ADN, mientras que R1 parece ser completamente dispensable ²⁹. Sin embargo, otros estudios han sugerido que R1 incrementa la estabilidad del complejo Myb-ADN, aunque esta estabilidad adicional parece ser parcialmente contrarrestada por la presencia de la región acídica N-terminal adyacente a R1. De acuerdo con estos estudios, R1 se une parcialmente a la hendidura mayor del ADN, pero la región N-terminal no permite una unión total. Los análisis de unión y protección sugieren que la función de la región ácida N-terminal es bloquear la unión de R1 al ADN ²⁸.

A pesar de que los efectos de R1 y la porción N-terminal no son grandes, pueden tener consecuencias importantes. En el caso de c-Myb, la especificidad de la secuencia de unión es relativamente baja, pudiendo unirse a múltiples secuencias, por lo que pequeños cambios en la afinidad de unión pueden afectar la ocupación relativa de los múltiples sitios de unión, lo que les permite ser usados para el control diferencial de unión a diferentes sitios. Más aún, la fosforilación de la región N-terminal por la caseína quinasa II inhibe completamente la unión entre c-Myb y sitios de poca afinidad, pero no es muy eficiente al inhibir la unión a sitios de alta afinidad. ²⁸.

El fragmento R1 se ha conservado a través de la evolución de los vertebrados en c-Myb, B-Myb y A-Myb, sugiriendo que tiene un papel importante en la función de estas proteínas. Estudios estructurales han confirmado que R1 se pliega formando tres hélices, al igual que R2 y R3; sin embargo, R1 no se encuentra presente en la mayoría de las proteínas relacionadas a Myb de los invertebrados ni en v-Myb. Ahora bien, estudios de mutagénesis y análisis estructurales han demostrado que remover R1 no afecta de forma dramática la actividad de unión al ADN de Myb en estudios *in vitro*. Esto deja dos posibles funciones para R1, que analizaremos de aquí en adelante. Primero, R1 puede funcionar como un dominio de unión “suplementario” al ADN, capaz de incrementar la afinidad de Myb por sitios específicos. Esto se supone porque Myb es capaz de unirse a un alto número de secuencias similares, y el factor común tiende a ser muy corto y poco específico pero la afinidad de Myb

por diferentes sitios puede variar sobre varios órdenes de magnitud. La extensión R1 podría hacer a las Myb más selectivas al reconocer secuencias adyacentes al núcleo del sitio de unión; por lo tanto, formas truncadas de Myb carentes de R1 serían más “promiscuas” de manera que podrían ser capaces de regular genes que c-Myb no puede. Un segundo posible rol para R1 es como un dominio de unión para una proteína específica; esto se propone como hipótesis ya que un número de proteínas con función de cofactores han sido identificadas, las cuales pueden unirse y modificar la actividad de proteínas relacionadas a Myb. Si R1 interactúa con una proteína inhibidora, su remoción haría versiones truncadas de Myb más activas, y les daría la posibilidad de transformar células¹⁹. A la fecha, no existe evidencia conclusiva que confirme alguna de estas hipótesis.

2.1.3. Secuencia de unión Myb-ADN:

El sitio de unión de las proteínas Myb al ADN ha sido estudiado también de manera profunda. En 1988, mediante pruebas de interacción de múltiples secuencias de ADN de gallina con v-Myb se determinó que la secuencia de unión de Myb a ADN se da mediante la interacción con la secuencia específica YAACKG, con Y siendo una pirimidina y K siendo guanina o timina³⁰. Estudios posteriores en 1991 identificaron la importancia de las bases vecinas, obteniendo una secuencia de unión de nueve bases YGRCVGTTR, donde Y es una pirimidina, R es una purina y V es cualquier base a excepción de timina. Esta secuencia se determinó mediante estudios de inmunoprecipitación de proteínas unidas a una secuencia de ADN³¹. Otros estudios determinaron que v-Myb y c-Myb tienen una preferencia *in vitro* por la secuencia de 8 pares de bases YAACKGHH, con H siendo Adenina, citosina o timina, y K siendo guanina o Timina.³²

Dada la importancia de los tres residuos de triptófano, se investigó si estos tienen interacción directa con el ADN. En 1990 se realizaron pruebas de unión de Myb al ADN con mutantes. Los resultados indicaron que no existía interacción directa entre los triptófanos con la cadena de ADN, y su principal rol es la estabilización de la estructura para permitir la interacción de los residuos vecinos con el ADN. Los estudios indicaron también que la eliminación del fragmento R1 no afecta la capacidad de unión del DBD al ADN.²⁹

En 1993 se reportó un estudio de la posible distribución de los repetidos R1, R2 y R3 sobre la secuencia de ADN objetivo²⁸. La secuencia usada durante el estudio fue 5' - CCTAACTGACACACATA-3' (Tabla 2), esta secuencia fue obtenida de los trabajos publicados por Nakagoshi *et al* en 1990, mediante sus estudios de afinidad de C-Myb con la secuencia del virus simio 40 (o papovirus-SV 40).

Tabla 2: sitios de unión de los tres fragmentos del dominio de unión de Myb

Objetivo	5' -CCTAACTGACACACATA-3'
Fragmento	
R1	5' -----ACATA-3'
R2	5' -----TGACAC-----3'
R3	5' ---TAACTG-----3'

Comparando los sitios de unión de las tres secciones, se observa que los tres poseen secuencias con un alto contenido de adenina, sin similitud aparente entre las tres (Tabla 2)

Estudios realizados en 1995 estudiaron las posibles causas de la contribución de R1 a la unión del DBD al ADN. Los estudios sugirieron que la presencia de aminoácidos ácidos en el primer giro de R1 genera una repulsión con el esqueleto de fosfatos de ADN, lo que conlleva a poca afinidad entre R1 y el ADN. En comparación, R2 y R3 poseen aminoácidos básicos en la región equivalente. Por otro lado, un cálculo de carga neta del repetido R1 generó un valor mucho menor a los valores calculados para R2 y R3, lo cual generaría una mayor dificultad para interactuar con el esqueleto de fosfatos del ADN.²⁷

Tabla 3: comparación de los aminoácidos en el primer giro de R1, R2 y R3 de m-Myb-C

Giro	-----1111-----2222-----
Hélice	-----////////////////-----////////////////-----////////////////-----
R1	LGKTRWTREEDEKLKLLVEQNGTDDWKVIANYLPNRTDVQCQHRWQKVLNPE
R2	LIKGPWTKEEDQRVIELVQKYGPKRWSVIAKHLKGRIGKQCRERWHNHLNPE
R3	VKKTTSWTEEDRIIYQAHKRLG-NRWAEIAKLLPGRTDNAIKNHNWSTMRKV

En la Tabla 3 se indican los aminoácidos en el primer giro a los que se le atribuye la capacidad de interacción entre el respectivo fragmento y el ADN. En R1 se tienen dos residuos de ácido aspártico consecutivos, mientras que en R2 se encuentran presentes un residuo de lisina y arginina, ambos aminoácidos básicos, y en R3 se presentan asparagina y arginina.

3. Hipótesis

Existe un núcleo hidrofóbico que es responsable de la estructura terciaria del repetido R1 de A-Myb.

Existe una región en el repetido R1 que permita su unión con la secuencia “canónica” de interacción con el ADN.

4. Objetivo general

Conocer la estructura del repetido R1 del dominio de unión al ADN de la A-Myb de humano

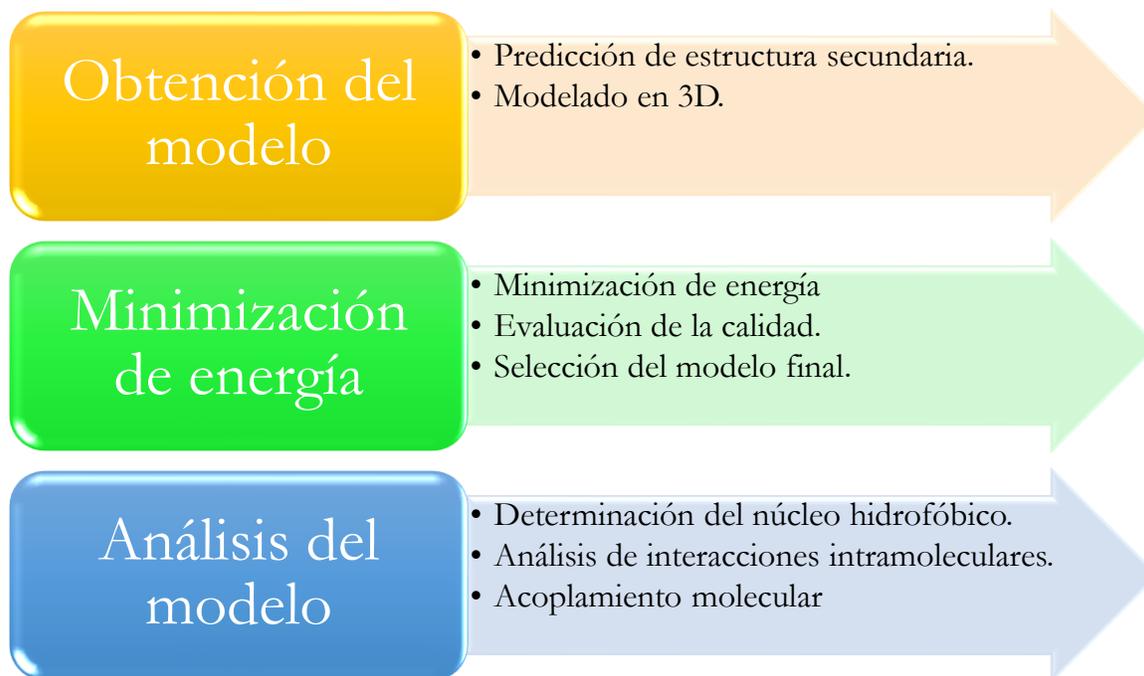
Objetivos particulares

- Obtener un modelo de la estructura del repetido R1 empleando métodos de modelado molecular.
- Determinar las interacciones que dan estabilidad a la estructura del modelo del repetido R1.
- Obtener experimentalmente el repetido R1 mediante métodos de expresión recombinante.
- Caracterizar el repetido R1 obtenido experimentalmente.

5. Metodología.

5.1. Modelado Molecular.

El trabajo de modelado se dividió en tres pasos:



5.1.1. Obtención del modelo.

Predicción de estructura secundaria y modelos iniciales

Se buscó la secuencia de R1 de A-Myb de humano para llevar a cabo las operaciones de predicción de estructura y el modelado molecular. La secuencia se muestra a continuación.

Tabla 4: Secuencia de R1 de A-Myb de humano

Nombre	A-Myb R1 [Homo sapiens]	
	20	40
Secuencia	LWNRVKWTRDEDDKLLKLLVEQHGTDDWTLIASHLQNRSDFCQHRWQKVLNPE	

Se seleccionaron cinco servidores en línea para llevar a cabo la predicción de la estructura secundaria y el modelado molecular. Los cinco servidores elegidos han reportado una alta exactitud en modelado y han obtenido un puntaje alto en los experimentos de Evaluación crítica de las técnicas para la predicción estructural proteica (CASP, del inglés *Critical Assessment of Techniques for Protein Structure Prediction*). También se observó que los cinco servidores tuvieran algoritmos de predicción variados.

Jpred4: Jpred4 funciona con el algoritmo JNet, algoritmo que emplea una serie de dos redes neuronales entrenadas para la predicción de diferentes tipos de estructuras secundarias a partir de una secuencia, y, de ser necesario, emplea una red para identificar la estructura secundaria más probable de entre las predichas. El servidor lleva a cabo múltiples alineamientos de secuencia en PSI-BLAST, así como alineamientos empleando el software HMMer 3 para la determinación de la predicción final. Este servidor tiene una exactitud auto reportada de hasta 82%.³³

PredictProtein: *PredictProtein* es uno de los servidores más antiguos dedicados a la predicción de estructura secundaria. Este servidor ofrece un rango de predicciones de estructura y función muchos de los cuales combinan aprendizaje automático con información evolucionaria, la cual ha sido mejorada con algoritmos para un alineamiento más rápido. *PredictProtein* determina la estructura secundaria de cada residuo, así como la accesibilidad del disolvente, desorden de la proteína, flexibilidad del esqueleto, búsqueda de puentes disulfuro, conservación de secuencia, y varias otras opciones para casos particulares. Emplea PSIBLAST para la búsqueda de homólogos y realiza también una simplificación de resultados, eliminando los homólogos menos probables.³⁴

RaptorX: funciona en casos donde PSI-BLAST no, ya que lleva a cabo la predicción de estructura secundaria y modelado de la estructura en 3D sin la necesidad de múltiples homólogos. RaptorX sobresale en el alineamiento de moléculas objetivo con una identidad de secuencia menor al 30% con estructuras resueltas en el PDB. El servidor toma en consideración el número de homólogos no redundantes disponibles para la secuencia problema y una estructura plantilla para evaluar la calidad de la información contenida en los

perfiles de secuencia. Se emplean campos neurales condicionales (CNF) para integrar una variedad de señales biológicas contextuales en una función de evaluación probabilística no lineal. Finalmente, el servidor Raptor X tiene implementado un procedimiento de hilado de múltiples moldes (MTT), que permite el uso de múltiples plantillas para modelar una única secuencia.^{35,36}

SWISS-MODEL: Es un servidor para el modelado comparativo automático de estructuras tridimensionales de proteínas. En línea desde 1993, es el servidor web libre para el modelado de estructuras más activo. Sus funciones de modelado incluyen el modelado de complejos homoméricos y heteroméricos a partir de las secuencias dadas para cada uno de los monómeros. Emplea además un motor de modelado ProMod3, que incrementa su exactitud al momento de generar los modelos, así como un método de estimación de la calidad basado en QMEAN. El servidor lleva a cabo el modelado mediante el empleo de estructuras de proteínas homologas que sirvan como plantilla, estas son identificadas mediante una búsqueda BLAST, para casos con alta homología, y HHblits, que funciona en casos de homología remota.^{37,38}

I-TASSER: (Iterative Threading ASSEmbly Refinement) es un método bioinformático para la predicción de estructuras secundarias y el modelado de la estructura terciaria de proteínas a partir de una secuencia. I-TASSER emplea un acercamiento jerárquico a la predicción de la estructura de proteínas, identificando primero moldes de estructuras en el PDB mediante el método de *threading* múltiple. El modelo final es construido mediante el ensamblaje de varios fragmentos obtenidos por simulaciones iterativas a partir de los diferentes moldes obtenidos en el paso de hilado.^{38,39}

5.1.2. Minimización de la energía y análisis y selección de modelos.

Minimización de energía.

Para la minimización de energía de los modelos obtenidos se empleó el servidor YASARA. Este servidor lleva a cabo simulaciones de dinámica molecular en disolvente explícito

empleando un método parcialmente basado en conocimiento de un campo de fuerza derivado de AMBER.⁴⁰

Evaluación de los modelos.

Los modelos moleculares obtenidos fueron evaluados utilizando el servidor Web MolProbity. Este es un servicio web para la validación de estructuras que provee una evaluación de amplio espectro de la calidad de un modelo a un nivel global y local para proteínas y ácidos nucleicos. Se basa en la capacidad y la sensibilidad obtenida mediante la optimización de la colocación de hidrógenos y el análisis del contacto de todos los átomos, complementado con las versiones más actuales de los criterios de geometría covalente y de ángulos de torsión. Correcciones locales al modelo pueden realizarse automáticamente en MolProbity y todo el diagnóstico es presentado en una tabla y en forma gráfica que sirven como guía para la reconstrucción manual del modelo.⁴¹. Durante la evaluación se emplearon las siguientes condiciones:

- Hidrógenos: Adición de hidrógenos mediante posiciones nucleares.
- Corrección automática de Asn/Gln/His recomendadas

Análisis visual comparativo con estructuras reportadas de R1 se llevaron a cabo empleando el software UCSF ChimeraX, desarrollado por el departamento de Recursos para la Biocomputación, Visualización e Informática de la Universidad de California, San Francisco.

Selección del modelo final

Se eligió el modelo de R1 con base en los resultados de las evaluaciones con MolProbity y ChimeraX,

5.1.3. Análisis del modelo

Determinación del núcleo hidrofóbico

Se evaluó la hidrofobicidad de la superficie del fragmento R1 empleando el Software UCSF ChimeraX. Posteriormente se calculó el área de superficie accesible relativa (RSA) para cada uno de los residuos de la proteína. El área de superficie accesible es el área de la superficie de una biomolécula que es accesible a un disolvente. Cuando se analiza el únicamente la exposición de un residuo en una biomolécula, se denomina área de superficie accesible del residuo. Esta última se encuentra relacionada con el arreglo espacial y el empaquetado del residuo. Con base en el valor del área de superficie accesible se puede clasificar a los residuos como “ocultos” o “expuestos”. para determinar si un residuo se encuentra o no expuesto es necesario conocer el área de superficie accesible del residuo, la cual es normalmente expresada como área de superficie accesible relativa (RSA). Esta es calculada al comparar el área accesible del residuo en la biomolécula con la valor máximo posible para ese residuo en un trímero Gly-X-Gly (ó Ala-X-Ala), donde X es el residuo de interés.⁴². Este cálculo se realizó empleando el servidor POLYVIEW-2D, el cual genera una secuencia de anotaciones que incluye el RSA para cada residuo, representado con valores en un rango de 0 a 9, siendo 0 completamente enterrado (0-9% RSA) y 9 totalmente expuesto (90-100%). Para continuar con la determinación del núcleo hidrofóbico, se establecieron valores límites de RSA para diferenciar los aminoácidos “ocultos” de los “expuestos”, Para este trabajo se clasificó como enterrado aquellos aminoácidos con un RSA menor al 20% (valores 0 y 1) ⁴²

Definiendo al núcleo hidrofóbico de una proteína hidrosoluble como el conjunto de residuos con baja accesibilidad del disolvente, los cuales se encuentran localizados en regiones con estructura secundaria regular y cuyas cadenas laterales no polares interactúan entre ellas⁴³, es posible considerar al conjunto de aminoácidos con cadenas laterales hidrofóbicas y con un bajo RSA como el núcleo hidrofóbico de nuestra proteína.

Análisis de interacciones intramoleculares

Para verificar que realmente exista interacción entre los aminoácidos seleccionados como parte del núcleo hidrofóbico, se utilizó el servidor *PIC: Protein Interaction Calculator*,⁴⁴. Este servidor considera que existen interacciones entre los residuos de Ala, Val, Leu, Met, Phe, Trp, Pro y Tyr si se encuentran dentro de un rango de 5 Å.

Para evaluar las interacciones entre todos los aminoácidos que conforman el fragmento R1, se utilizó el servidor web *INTAA*. Este servidor permite calcular el residuo de la matriz de la Energía de Interacción para la estructura de una proteína y además de proveer un análisis comprensivo en el caso de las interfases en complejos de Proteína-ADN. Esta aplicación tiene como objetivo ayudar a identificar residuos clave dentro de la estructura proteica que contribuyan significativamente a la estabilidad de la proteína.⁴⁵

Acoplamiento molecular

Para determinar si es posible la interacción entre la molécula de R1 modelada y el ADN se llevó a cabo un proceso de acoplamiento molecular, para ello se emplearon los siguientes servidores:

HDOCK: Posee un modelo híbrido de acoplamiento molecular, es decir, lleva a cabo un acoplamiento basado en plantillas y un acoplamiento libre, donde los casos con plantillas erróneas puedan ser verificadas con el protocolo de acoplamiento libre. El proceso del servidor consta de cuatro pasos: la recepción de las estructuras a acoplar, la búsqueda de homólogos en el Banco de Datos de Proteínas (PDB - Protein Data Bank) para cada una de las estructuras, la determinación si existen entradas en el PDB que incluyan homólogos de ambas moléculas y construcción de posibles estructuras de acuerdo con las plantillas, y finalmente un acoplamiento global tradicional. El algoritmo de acoplamiento emplea un programa propio llamado HDOCKlite, es empleado para muestrear globalmente posibles modos de unión en el servidor HDOCK. Se emplea además una nueva función de puntaje, la cual toma en consideración no solo las contribuciones de la red de la molécula blanco en la zona de interacción e inmediatamente vecinas, sino también de otras regiones en un radio

determinado por la distancia entre el receptor y el ligando. Para el muestreo rotacional se emplea un intervalo de 15° y se adopta una distancia de 1.2 \AA para la búsqueda traslacional. Para cada rotación se optimizan las 10 mejores traslaciones para su posterior búsqueda en las funciones de puntaje. Se conserva un modo de unión, que corresponde a la mejor traslación, para cada rotación. Estos modos de unión son ordenados y agrupados con base en su RMSD (root-mean-square deviation, raíz del error cuadrático medio), manteniéndose aquellos con mejor RMSD para cada modo de unión. Finalmente, el servidor genera un modelo para los 100 modos de unión con mejor puntaje.⁴⁶

HADDOCK: HADDOCK es un método de acoplamiento molecular llevado a cabo mediante conocimiento experimental sobre la región de interfaz entre las moléculas acopladas y/o su relación relativa. En los casos cuando no exista información experimental reportada o esta sea escasa, pueden ser usadas predicciones bioinformáticas de la interfaz. El servidor HADDOCK hace uso de datos de interacción fisicoquímica o biofísica al introducirlos como Restricciones de Interacciones Ambiguas para mediar y dirigir el proceso de acoplamiento molecular. HADDOCK automáticamente genera la topología de las moléculas que van a ser acopladas. El protocolo de acoplamiento consiste en tres pasos, que son una minimización de la energía de cuerpo-rígido, un afinamiento semiflexible en ángulos de torsión y un afinamiento final en disolvente explícito. Tras cada uno de estos pasos, las estructuras son evaluadas y ordenadas y aquellas con mejor estructura pasan al siguiente paso. La evaluación consiste en una suma ponderada de energías de Van der Waals, electrostáticas, solvatación y las restricciones en conjunto con la el área de superficie no accesible al disolvente.⁴⁷

NPDock: NPDock (Nucleic acid-Protein Dock) es un servidor web para el modelado de estructuras de complejos ARN-Proteína y ADN-Proteína. El servidor emplea el programa GRAMM (Global Range Molecular Matching) para generar estructuras geoméricamente plausibles de complejos proteína-ácido nucleico. Posteriormente se aplican las restricciones listadas por el usuario. Las posibles estructuras son evaluadas mediante un método diseñado para este tipo de complejos. Los complejos con la mejor puntuación son agrupados y se selecciona un modelo representativo de cada uno de los tres grupos más grandes. Finalmente,

se lleva a cabo un afinamiento local para optimizar las interacciones, y las estructuras resultantes son presentadas al usuario.⁴⁸

Molécula Blanco

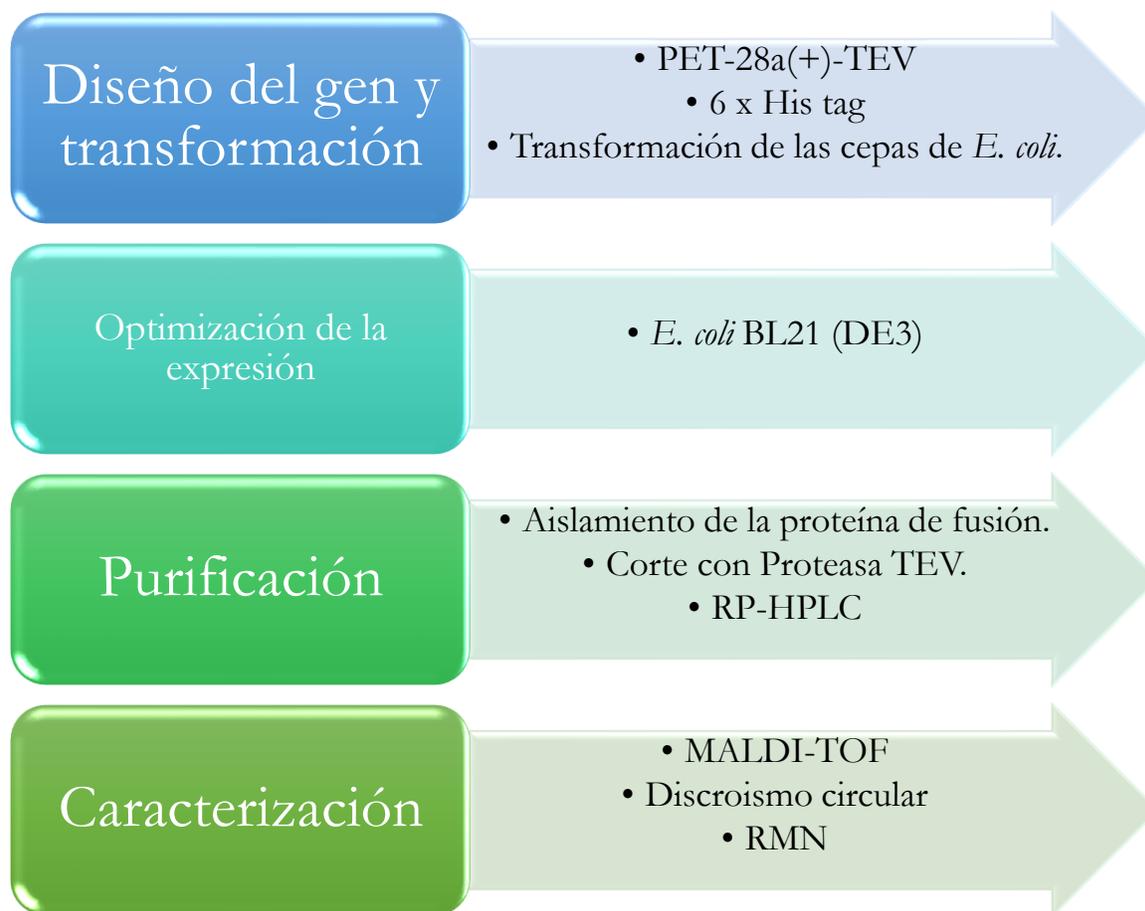
El blanco objetivo para la molécula modelada fue la molécula de ADN cuya secuencia se muestra en la tabla 5 y fue obtenida del PDB 1BDZ. Esta secuencia consta de 14 pares de bases que incluyen un sitio de unión de C-Myb de tipo YAACKG, esta secuencia fue la primera secuencia de unión de proteínas Myb determinadas y se considera como una secuencia de unión “canónica” para proteínas Myb. La molécula objetivo tiene además una cadena complementaria simétrica, la cual contiene el mismo sitio de unión en sentido contrario. Ambos sitios de unión se señalan en la tabla 5

Tabla 5: Secuencia “canónica” usada para las pruebas de acoplamiento molecular.

Secuencia	5' -ACAACTGCAGTTGT-3'
Complementaria	3' -TGTTGACGTCAACA-5'

5.2. Obtención de R1

Para traducir el enfoque teórico planteado inicialmente, se propuso un trabajo experimental, el cual se dividió en cuatro etapas, las cuales se representan en la siguiente imagen:



5.2.1. Diseño del gen y transformación

El gen se diseñó teniendo en cuenta que habría de expresarse y obtenerse en las cepas de *E. coli* Rosetta y BL21, de modo que alguno de estos sistemas permitiera obtener la proteína de fusión de forma soluble; para ello se utilizó el plásmido PET-28a(+)-TEV. La secuencia de la proteína de fusión se muestra en la Figura 7. El esquema muestra las tres regiones que conforman dicha proteína: la etiqueta de 6 histidinas empleada en la purificación (verde), el

sitio de corte con proteasa TEV (rosa) y la secuencia de aminoácidos del dominio R1 de hA-Myb (celeste).



Figura 7. Representación de la proteína de fusión (hR1A).

La transformación se llevó a cabo mediante choque térmico empleando células competentes de las cepas Rosetta y BL21 (DE3) de *E. coli* mediante el siguiente protocolo:

- I. Un volumen de 3 μ L de plásmido fue añadido a 50 μ L de células competentes de *E. coli* en un tubo de microcentrífuga y se mantuvieron en un baño de hielo por 30 min.
- II. El tubo se transfirió a un baño de agua a 42 °C por 45 s
- III. Se regresó el tubo al baño de hielo por 2 min más.
- IV. Se añadieron 450 μ L de medio Luria Bertani (LB) y se colocaron en incubación por 1 h.
- V. Las células obtenidas fueron sembradas en una caja con medio LB sólido con los antibióticos apropiados para cada cepa e incubadas por 14 h a 37 °C.
- VI. Una vez transcurrido este tiempo, se cuantificaron el número de colonias obtenidas

5.2.2. Optimización de la expresión.

A continuación, se estandarizaron las condiciones de expresión de la proteína de fusión para determinar las condiciones que permitieran obtener la mayor cantidad de proteína. Se eligieron los siguientes parámetros:

- Cepa de *E. coli*.
- Temperatura de expresión.
- Tiempo de expresión.

Para determinar qué cepa es la más adecuada para la expresión de R1 el gen se introdujo en las bacterias mediante la transformación de células competentes de las cepas Rosetta™ y BL21(DE3)™ de *E. coli*. A continuación, cada cepa transformada se inoculó en 10 mL de medio de cultivo Luria Bertani (LB) con los antibióticos apropiados (cloranfenicol 34 mg/mL

y kanamicina 30 mg/mL para Rosetta™ y kanamicina 30 mg/mL para BL21(DE3)™). Los 10 mL de medio se incubaron por 16 h a 37 °C, tras este periodo, se transfirieron 1.66 mL de medio inoculado a un matraz que contenía 50 mL de medio LB y los antibióticos respectivos (Proporción de preinóculo/medio de expresión: 1/30). Los matraces se incubaron a 37 °C monitoreando el valor de la densidad óptica a 600 nm. Cuando el cultivo alcanzó una densidad óptica de 0.6 se agregó el agente inductor de la expresión: isopropil β-tiogalactopiranos (IPTG) utilizando una concentración final en el cultivo bacteriano de 0.5 mM.

La expresión de la proteína de fusión fue monitoreada en ambas cepas, para ello se tomaron muestras a un tiempo de expresión de 2 h, 4 h, 6 h, 8 h, 12 h y 24 h. Estas muestras fueron centrifugadas, recolectando únicamente las células, las cuales se resuspendieron en una solución amortiguadora de fosfato salino (PBS) a pH 6.5 antes de someterse a lisis con ultrasonido. Tras la lisis, las muestras fueron nuevamente centrifugadas, recolectando las fracciones solubles e insolubles por separado. Se realizó electroforesis en gel de poliacrilamida con SDS (SDS-PAGE) para ambas fracciones de las muestras, monitoreando la región indicada por el marcador de peso molecular cercana a 8 kD.

Para la determinación de la temperatura de expresión se repitió el proceso anterior cambiando la temperatura tras la inducción con IPTG a 30 °C.

Tomando los resultados de las pruebas anteriores, se repitió nuevamente la expresión de la proteína, tomando muestras en los siguientes tiempos tras la inducción de la expresión: 2 h, 4h, 6 h, 8 h, 12 h y 24 h. Posteriormente, las proteínas se separaron mediante SDS-PAGE en un gel de poliacrilamida a pH 6.8 para determinar el tiempo en que se obtenía la mayor expresión de la proteína.

6. Resultados y análisis de resultados

6.1. Modelado molecular

6.1.1. Predicción de estructura secundaria y modelado molecular

Los resultados de predicción de estructura secundaria y modelado de cada uno de los servidores utilizados se resumen a continuación:

Jpred4: Los resultados de predicción de estructura secundaria se muestran en la Tabla 6 y coinciden con la información sobre la estructura de las secuencias repetidas del dominio de unión Myb, las cuales presentan el motivo hélice-giro-hélice. De acuerdo con este cálculo la estructura presenta tres hélices, indicadas en la Tabla 6 como **H**, entre los residuos Arg9-His22, Trp27-His33, y Asp39-Gln47. Si bien los dos métodos empleados por el servidor para determinar la estructura (JnetHMM y JnetPSSM) presentaron ciertas diferencias en el rango de la tercera hélice, la red jurado determinó la estructura secundaria en JnetPred como la más probable.

Tabla 6: Resultados obtenidos con Jpred4

Secuencia	LWNRVKWTRDEDDKLLKLVQHGTDWTLIASHLQNRSDFCQHRWQKVLNPE
JnetPred	-----HHHHHHHHHHHHHHHH-----HHHHHHH-----HHHHHHHHHH-----
JnetHMM	-----HHHHHHHHHHHHHHHH-----HHHHHHH-----HHHHHHHHHH-----
JnetPSSM	-----HHHHHHHHHHHHHHHH-----HHHHHHH-----HHHHHHHH-----
Jnet jury	-----*-----**-----
Jnet Rel	9998876278999999999998606874688887526887128877644078999
JnetSol25	-B-----B---BB-BBB-BB--BB--B--BB--B--BB--BBB-BB--BB---
Sol15	-----B--B--B--BB--B-----B--B-----
Sol10	-----B-----B-----B-----

H: aminoácidos que forman parte de una hélice α ; **B:** Aminoácidos ocultos (Buried). Los números indican la confiabilidad de los resultados de predicción.

Se observa, además, que la confiabilidad de la predicción de la estructura (JnetRel) es alta tanto en las zonas abarcadas por las hélices como en los giros y zonas de inicio y final, presentando solo poca confianza en los aminoácidos iniciales y finales de cada hélice. Otro resultado a mencionar es la presencia de aminoácidos con poca accesibilidad al disolvente, reportada en las filas como JnetSol25, JnetSol15 y JnetSol10, indicando una accesibilidad

del 25%, 15% y 10% respectivamente. Se calculó que en cada hélice existe al menos un aminoácido con una accesibilidad menor al 10%, que muy posiblemente forme parte del núcleo hidrofóbico que da estabilidad a la estructura.

PredictProtein: Estos resultados muestran presencia de hélices entre los aminoácidos Arg9-His22, Trp27-His33, Asp39-Gln47. Estos resultados coinciden con la estructura secundaria predicha por Jpred4. Por otro lado, PredictProtein predijo también las características de accesibilidad del disolvente a cada aminoácido de la proteína, se observa un cierto grado de coincidencia entre estos resultados con los resultados de JPred4 para una accesibilidad del 10%.

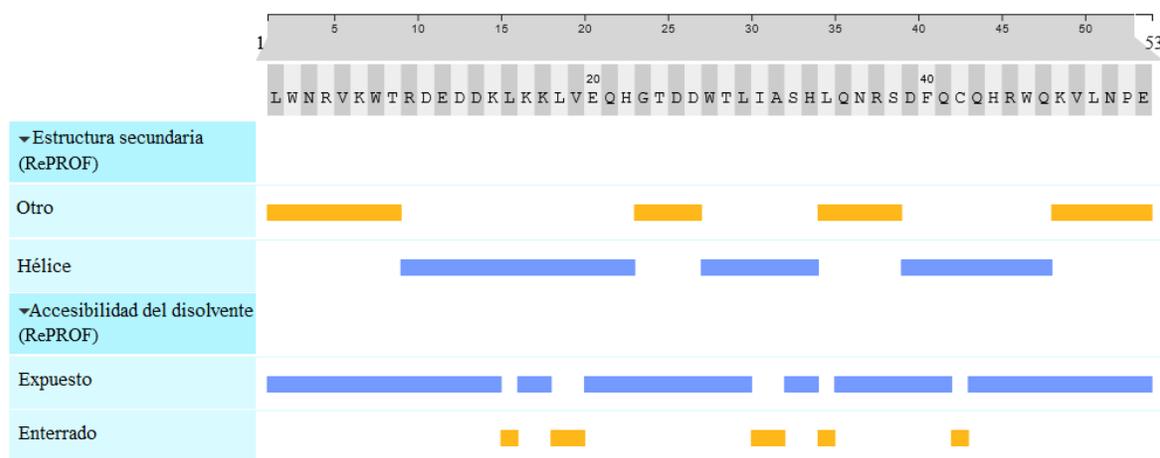


Figura 8. Predicción de estructuras secundaria del repetido R1 del DBD de A-Myb obtenidos por PredictProtein.

RaptorX: RaptorX calculó la presencia de tres hélices en los siguientes intervalos: Arg9-His22, Trp27-His33, Asp39-Val49. Estos resultados coinciden con las predicciones obtenidas por JPred 4 y PredictProtein para las primeras dos hélices, pero difieren para los obtenidos en la tercera hélice.

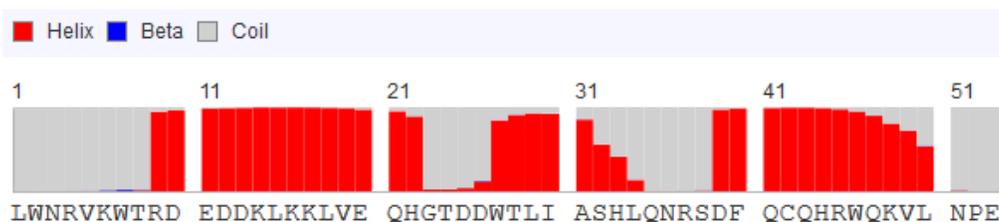


Figura 9 Predicción de estructuras secundaria del repetido R1 del DBD de A-Myb obtenidos por RaptorX. Las barras rojas indican la probabilidad de la presencia de una hélice α

reporta la estructura de un complejo de ADN con un dominio de unión de C-Myb (R1R2R3). Es importante mencionar que la posición de las tres hélices en el modelo difiere de la posición indicada en la Tabla anterior. La tercera hélice se ubica, en el modelo, entre los aminoácidos Asp39-Val49, siendo el modelo similar al elaborado por el servidor RaptorX.

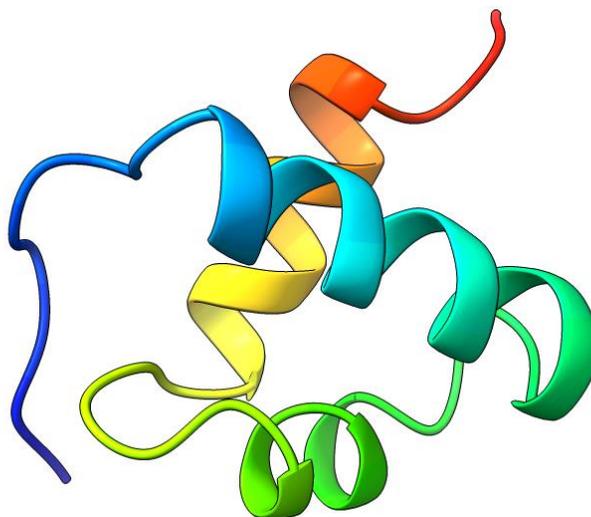


Figura 11. Modelo tridimensional del repetido R1 del DBD de A-Myb generado por I-TASSER

SWISS-MODEL: Para llevar a cabo el modelado se empleó como molde el PDB 1mbe.1.A: *MYB PROTO-ONCOGENE PROTEIN*, el cual es la estructura del fragmento R1 de C-Myb de *Mus musculus* (mR1C) y posee una identidad de secuencia del 71.15% con hR1A. El alineamiento de esta secuencia con la secuencia hR1A se muestra en la Figura 12.

```

Model_01 L WNRVKWT RDEDDKLLKLV EOHGTDD WTLIASH LQNRSD FQCQHRWQKVLNPE 53
1mbe.1.A - LGKTRWT REEDEKLLKLV EONGTDD WKVIAN YLPNRT DVQCQHRWQKVLNPE 52
  
```

Figura 12. Alineamiento generado por SWISS-MODEL. Los recuadros representan las hélices de la estructura

SWISS-MODEL también predijo la presencia de tres hélices alfa; sin embargo, SWISS-MODEL predijo una hélice H1 menor en longitud a los tres modelos anteriores, la segunda hélice coincide en todos los casos y la tercera hélice es igual a la predicha por JPred4 y diferente a la estimada por PredictProtein, RaptorX y I-TASSER.

Los estimadores de la calidad global del modelo fueron generalmente favorables. Estos estimadores son los valores de los parámetros GMQE y QMEAN. El valor de GMQE es un

estimado de la calidad que combina las propiedades del alineamiento de secuencias con la estructura molde. Puede tener un valor de 0 a 1, siendo un número mayor un indicador de mayor precisión al momento de construir un modelo empleando ese alineamiento. Para el caso de hR1A con el PDB 1mbe . 1 . A, se obtuvo un valor de GMQE de 0.76, indicando una alta precisión al construir el modelo empleando 1mbe . 1 . A como plantilla.

El valor de QMEAN es un estimador compuesto basado en las diferencias geométricas que provee un estimado absoluto de calidad global y local. El valor Z-score de QMEAN indica si el valor de QMEAN del modelo es comparable al valor que se esperaría de estructuras de tamaño similar determinadas experimentalmente. Valores cercanos a cero indican una concordancia adecuada entre el modelo y las estructuras experimentales. El valor de QMEAN para el modelo de hR1A fue de -0.63, indicando una buena concordancia entre este modelo y estructuras de tamaños similares.

La Figura 13 desglosa los cuatro componentes del valor de QMEAN global para el modelo. En este gráfico, las áreas en color blanco indican que la propiedad es similar a la que se esperaría de estructuras experimentales. La Figura 14, por su parte, muestra los valores estimados de la calidad local del modelo, siendo residuos con valores inferiores a 0.6, residuos con baja calidad. Para nuestro modelo, las únicas zonas de la estructura con baja calidad fueron las secuencias iniciales y finales, que corresponden a zonas desordenadas. Finalmente, el modelo generado por SWISS-MODEL se muestra en la Figura 15.

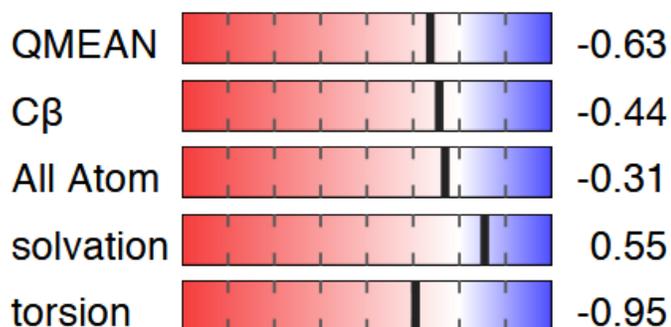


Figura 13. Estimador QMEAN global del modelo

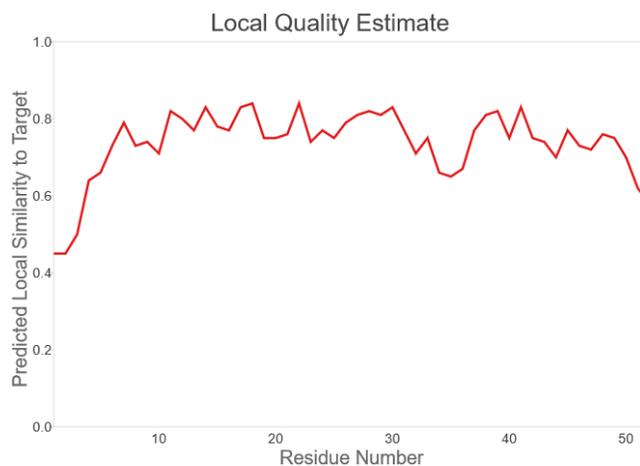


Figura 14. Estimado de la calidad para cada residuo del modelo.

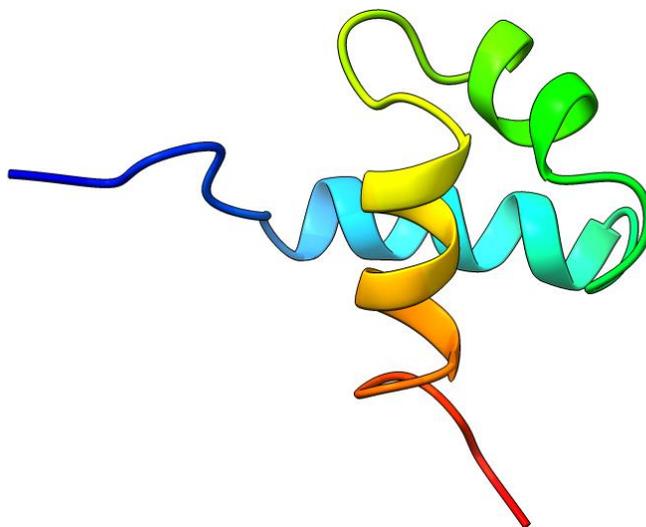


Figura 15. Estructura modelada por SWISS-MODEL

La siguiente tabla resume los resultados de predicción de los elementos de estructura secundaria de nuestra proteína. Todos los métodos predijeron la presencia de tres hélices, variando únicamente en la posición de ésta.

Tabla 8: Resumen de los resultados de predicción de elementos de estructura secundaria

Método	H1	H2	H3
JPred4	Arg9-His22	Trp27-His33	Asp39-Gln47
PredictProtein	Arg9-His22	Trp27-His33	Asp39-Gln47
RaptorX	Arg9-His22	Trp27-His33	Asp39- Val49
I-TASSER	Arg9-His22	Trp27-His33	Asp39-Gln47
SWISS-MODEL	Arg9-Gln21	Trp27-His33	Asp39-Gln47

Comparación gráfica de los modelos.

Los resultados se muestran en la siguiente serie de imágenes de la Figura 16. En naranja se muestran los resultados obtenidos por RaptorX, en verde los obtenidos mediante I-TASSER y en cian los obtenidos por SWISS-MODEL.

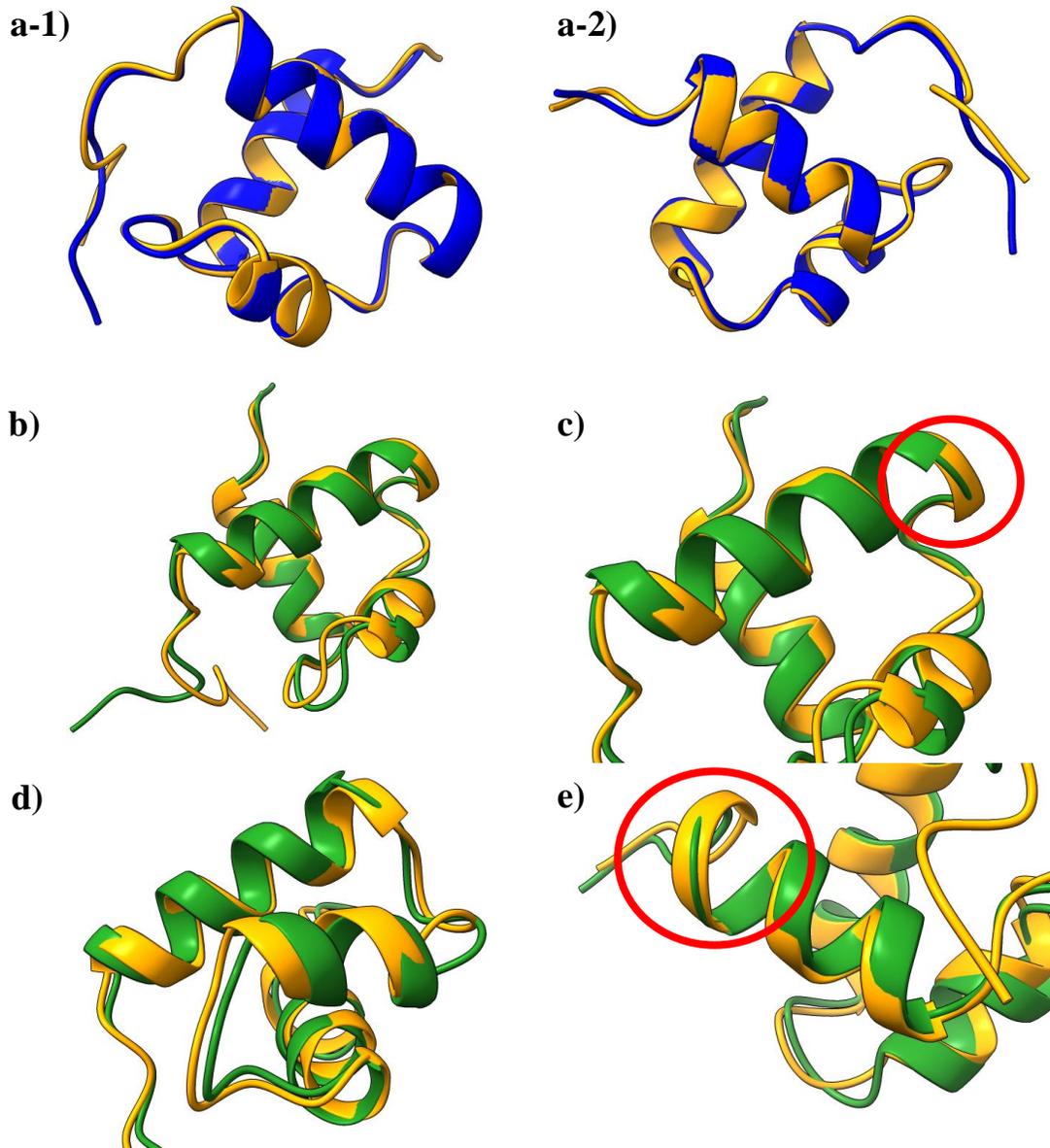


Figura 16. Comparación de los tres modelos del repetido R1 de la proteína A-Myobobtenidos: en naranja se muestra el modelo obtenido por RaptorX, en azul marino, el obtenido por I-TASSER y en verde se muestra el obtenido por SWISS-MODEL. a-1) y a-2) Superposición de los modelos de RaptorX y I-TASSER, ambos modelos son muy similares; b) Superposición de los modelos de RaptorX y SWISS-MODEL, es posible apreciar la diferencia entre los modelos; c) Se observa la hélice H1, la cual tiene una longitud mayor en el modelo de RaptorX, d) la hélice H2 se encuentra en la misma posición en ambos modelos, la diferencia observada en la superposición deriva de las diferencias entre las H1 y H3 de cada modelo, e) la hélice H3 tiene mayor longitud en el modelo de RaptorX.

6.1.2. Minimización de la energía.

Tras la minimización de la energía mediante el servidor YASARA, se obtuvieron tres nuevos sistemas de coordenadas o modelos, uno para cada uno de los suministrados. Los resultados de la evaluación de estos modelos mediante *MolProbity* se resumen en las tablas 9, 10 y 11.

Tabla 9: Evaluación del modelo obtenido por RaptorX: R1_RaptorX-1_YAS

Contactos	Clashscore	0		
	Clashscore: traslapamientos estéricos (> 0.4 Å) por cada 1000 átomos			
Geometría de la proteína	Rotámeros pobres	1	1.96%	<0.3%
	Rotámeros favorecidos	48	94.12%	>98%
	Ramachandran atípicos	0	0.00%	<0.05%
	Ramachandran favorecidos	50	98.04%	>98%
	Z-score	-0.76 ± 1.09		< 2
	Cβ desviaciones >0.25 Å	0	0.00%	0
	Enlaces erróneos	1/475	0.21%	0%
	Ángulos erróneos	1/643	0.16%	<0.1%

Tabla 10: Evaluación del modelo obtenido por I-TASSER: R1_I-TASSER-1_YAS

Contactos	Clashscore	0		
	Clashscore: traslapamientos estéricos (> 0.4 Å) por cada 1000 átomos			
Geometría de la proteína	Rotámeros pobres	0	0.00%	<0.3%
	Rotámeros favorecidos	50	98.04%	>98%
	Ramachandran atípicos	0	0.00%	<0.05%
	Ramachandran favorecidos	48	94.12%	>98%
	Z-score	-0.94 ± 1.11		< 2
	Cβ desviaciones >0.25 Å	0	0.00%	0
	Enlaces erróneos	1/475	0.21%	0%
	Ángulos erróneos	1/643	0.16%	<0.1%

El principal parámetro calculado por MolProbity en el análisis de contactos es la *Clashscore*, la cual es el número de choques serios en la estructura por cada 1000 átomos. Un choque serio ocurre cuando dos átomos que no tienen una interacción de tipo donador-aceptor se traslapan espacialmente por más de 0.4 Å. Traslapos tan grandes no pueden existir en la molécula real, por lo que su presencia indica que al menos uno de los dos átomos se encuentra

modelado incorrectamente. Como puede observarse, los tres modelos poseen una Clashscore de 0, lo cual es indicativo de una buena calidad.

Tabla 11: Evaluación del modelo obtenido por SWISS-MODEL: RI_SM-1_YAS

Contactos	Clashscore	0		
	Clashscore: traslapamientos estéricos (> 0.4 Å) por cada 1000 átomos			
Geometría de la proteína	Rotámeros pobres	1	2.00%	<0.3%
	Rotámeros favorecidos	49	98.00%	>98%
	Ramachandran atípicos	1	2.00%	<0.05%
	Ramachandran favorecidos	43	86.00%	>98%
	Z-score	-3.62 ± 0.94		< 2
	C β desviaciones >0.25 Å	0	0.00%	0
	Enlaces erróneos	1/467	0.21%	0%
	Ángulos erróneos	1/632	0.16%	<0.1%

El análisis de los rotámeros hace referencia a la geometría de las cadenas laterales de los aminoácidos. El número indica el número de aminoácidos en la categoría de pobres o favorecidos y el porcentaje total de residuos que cae en esa categoría. De los tres modelos, el modelo obtenido por I-TASSER es el que obtuvo los mejores valores en este análisis. El análisis de Ramachandran muestra un número de aminoácidos con ángulos phi o psi pobres o favorecidos. En este caso, el modelo obtenido por RaptorX presenta los mejores valores. La Z-score compara la distribución de Ramachandran obtenida contra la distribución esperada. En este caso tanto RaptorX como I-TASSER generaron mejores valores que el modelo obtenido mediante SWISS-MODEL.

A primera vista, el análisis de la geometría de la proteína parece favorecer los modelos obtenidos por I-TASSER y RaptorX, al estos presentar la mayor cantidad de parámetros dentro de los valores aceptables. Sin embargo, para este caso particular de hR1A, el análisis de la geometría parece ser poco útil al momento de elegir un modelo en particular. Debido a la corta longitud de hR1A, la presencia de un único residuo con geometría fuera de los límites establecidos por MolProbity ocasiona que la estructura falle en los parámetros evaluados por el servidor.

Selección del modelo final.

Se compararon visualmente los modelos obtenidos tras la minimización de la energía con la estructura reportada en el PDB 1MBE se resume en la Figura 17. Se observa que todos los modelos obtenidos presentan una alta similitud con la estructura determinada experimentalmente, mostrando las variaciones más notables en las zonas desordenadas de la proteína. La mayor alineación se observa con el modelo obtenido mediante SWISS-MODEL, esto se debe a que SWISS-MODEL empleó al PDB 1mbe como plantilla para la elaboración del modelo.

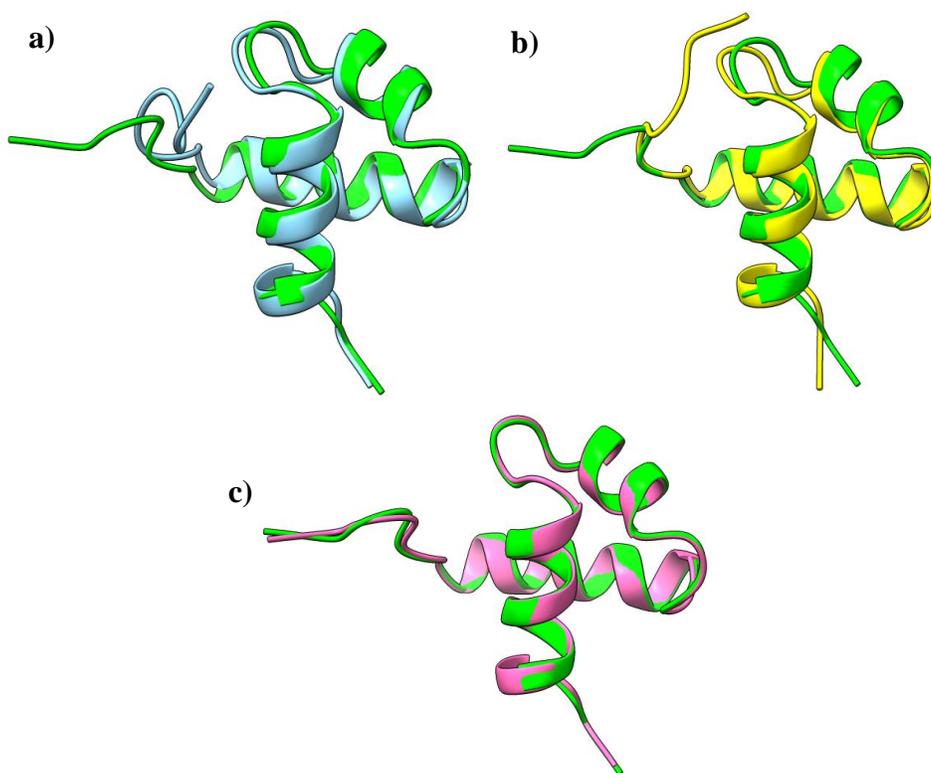


Figura 17. Comparación de la estructura reportada en el PDB con código de entrada 1MBE1.A (color verde), que corresponde al dominio de R1 de C-Myb de ratón con: a) el modelo obtenido mediante RaptorX (azul), b) el modelo obtenido mediante I-TASSER (amarillo), y c) el modelo obtenido mediante SWISS-MODEL (rosa).

Dado la importancia que se le ha dado a los tres triptófanos de los repetidos de Myb para la estabilidad estructural, se compararon visualmente la posición de estos en los modelos con la estructura experimental, los resultados se muestran en la Figura 18. Se puede observar que la mayoría de los triptófanos en los modelos se encuentran en posiciones cercanas a las de los triptófanos de 1MBE; sin embargo, en el caso de los modelos de RaptorX y I-TASSER

los rotámeros de dos de los tres residuos eran diferentes, en el caso del modelo de SWISS-MODEL, solo uno de los residuos difería de 1MBE. La orientación de estos rotámeros de triptófano puede afectar la estructura del núcleo hidrofóbico de la proteína y la estabilidad de la estructura en general.

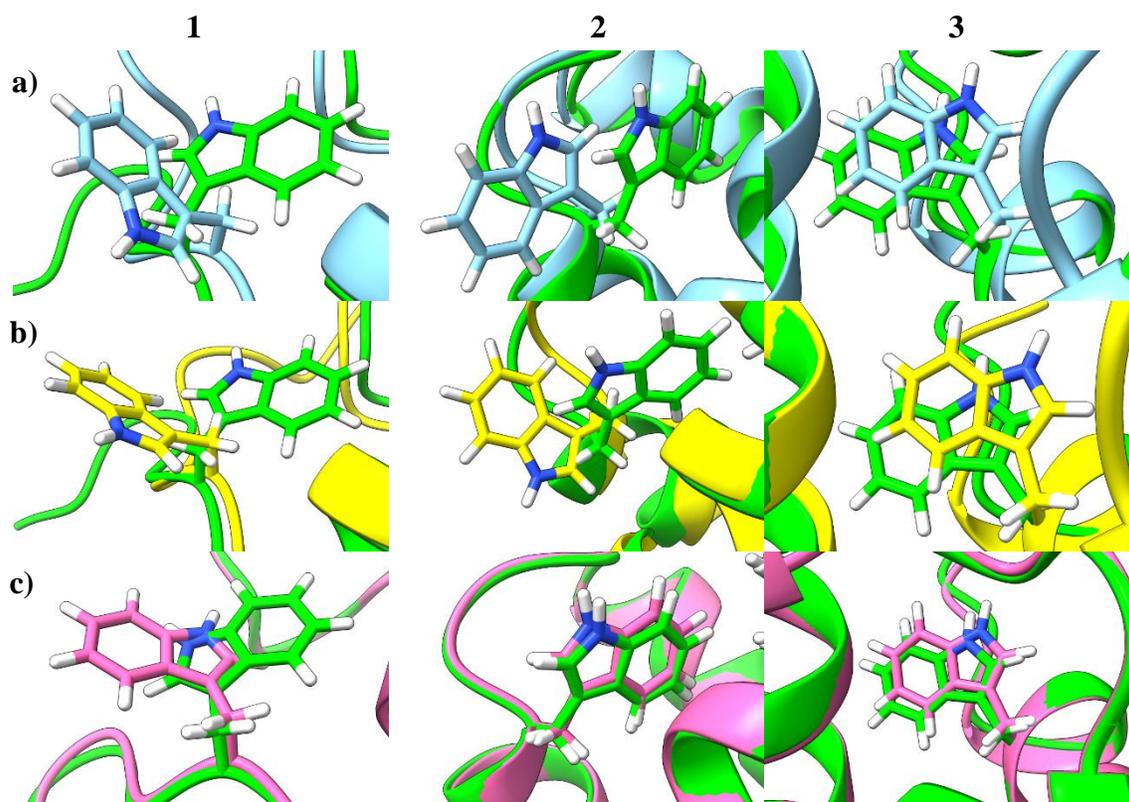


Figura 18. Comparación de la posición espacial de los residuos de triptófanos conservados entre la estructura de R1 reportada en el PDB 1MBE (color verde) y: a) el modelo obtenido mediante RaptorX (azul), b) el modelo obtenido mediante I-TASSER (amarillo), y c) el modelo obtenido mediante SWISS-MODEL (rosa). La columna 1 representa al primer triptófano, la columna 2 representa al segundo triptófano, y la columna 3 al tercer triptófano.

Es notable que el residuo que presentó diferencias en los tres modelos es el residuo TRP7, el cual se encuentra al principio de la estructura en una zona sin elementos de estructura secundaria, presentando el mayor número de grados de libertad al no tener muchas restricciones espaciales.

Empleando el Software USCF ChimeraX se determinó la hidrofobicidad de la superficie de cada uno de los tres modelos. Este software emplea la escala Kyte-Doolittle la cual asigna valores de Kd de hidrofobicidad a cada aminoácido, siendo los aminoácidos hidrofóbicos

asignados valores positivos los cuales disminuyen al disminuir la hidrofobicidad asignando valores negativos a los aminoácidos hidrofílicos.⁴⁹ Los resultados se resumen en la Tabla 12.

Tabla 12: Hidrofobicidad promedio de superficie para cada modelo

MODELO	HIDROFOBICIDAD
R1_RAPTORX-1_YAS	-3.309
R1_I-TASSER-1_YAS	-3.077
R1_SWISS_MODEL-1_YAS	-4.32

Se observa el modelo obtenido por SWISS-MODEL contiene la superficie menos hidrofóbica; esto significa que la mayoría de los residuos expuestos al disolvente son hidrofílicos lo cual produce un mayor empaquetamiento de los aminoácidos hidrofóbicos en el núcleo de nuestra estructura.

Considerando los puntos de similitud con el PDB 1MBE y el empaquetamiento de los modelos, se decidió continuar los análisis empleando únicamente el modelo para hR1A obtenido mediante SWISS-MODEL: R1_SWISS-MODEL_1_YAS.pdb

6.1.3. Análisis del modelo

Hidrofobicidad de la superficie.

En la Figura 19 y Figura 20 se muestran una serie de imágenes generadas de la hidrofobicidad de la superficie de la proteína, pueden notarse un mayor número de regiones hidrofílicas y neutras que de regiones hidrofóbicas. La región hidrofóbica de mayor superficie corresponde a la zona desordenada N-terminal

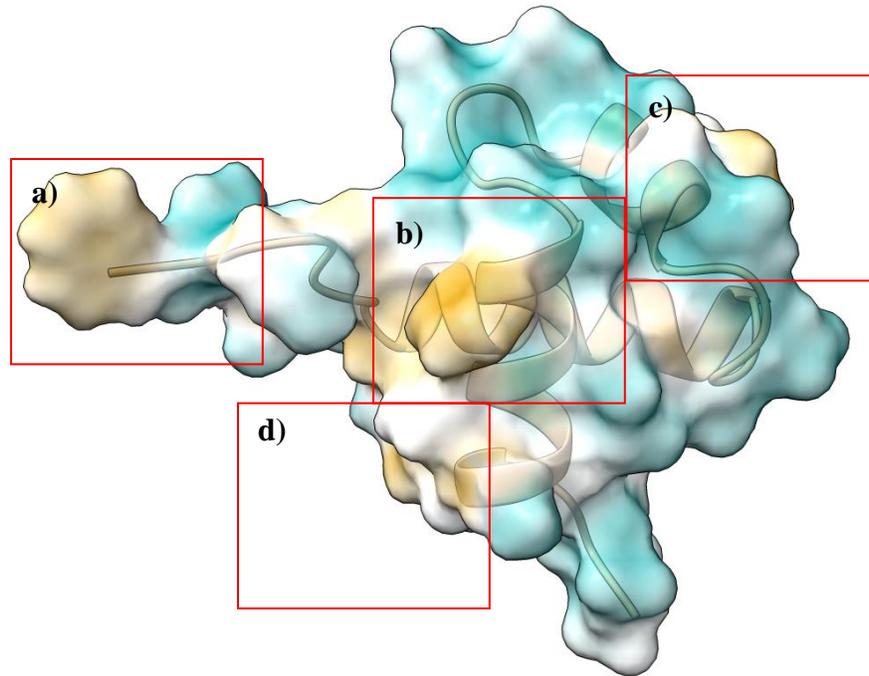


Figura 19. Análisis de hidrofobicidad de la superficie del repetido R1 de la proteína A-Myb. Las zonas de menor hidrofobicidad se representan en color azul y las de mayor en naranja. Se indican las zonas de mayor hidrofobicidad visibles.

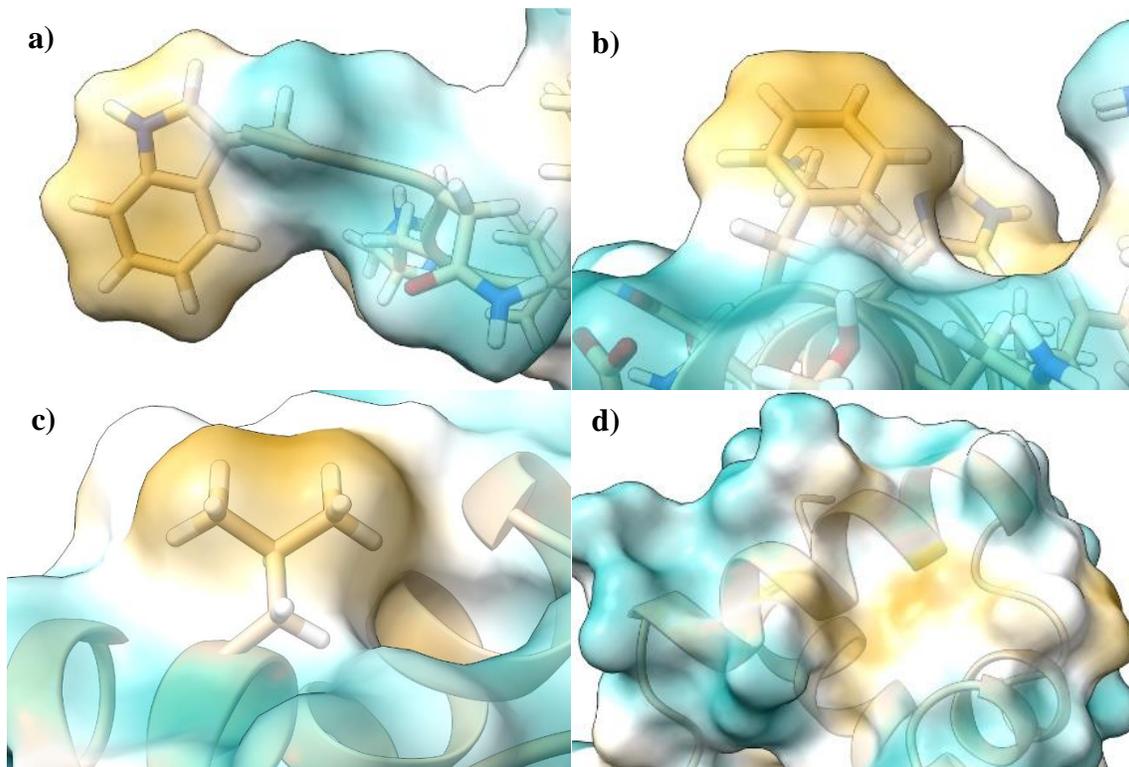


Figura 20. Acercamiento a las regiones de mayor hidrofobicidad indicados en la Figura 19: a) TRP2 en la región desordenada N-terminal, b) PHE40 en la parte expuesta de la tercera hélice; c) LEU29 en la segunda hélice, d) región entre las hélices.

Accesibilidad de disolvente y núcleo hidrofóbico

Los resultados de los cálculos de RSA realizados con POLYVIEW-2D se muestran en la Figura 21. Se muestra el valor de RSA para cada residuo, representado con valores en un intervalo de 0 a 9, siendo 0 completamente enterrado (0-9% RSA) y 9 totalmente expuestos (90-100%). Para este trabajo se clasificó como enterrado aquellos aminoácidos con un RSA menor al 20% (valores 0 y 1).

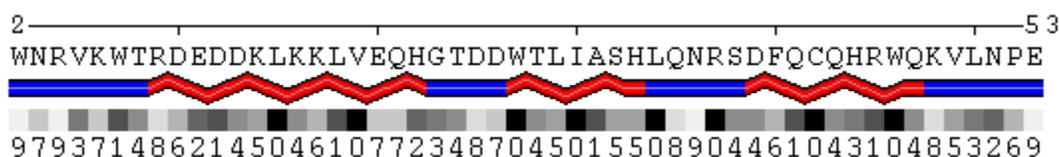


Figura 21. Área de superficie accesible relativa de cada residuo.

Los resultados obtenidos con POLYVIEW-2D indican que existen 14 aminoácidos con un RSA menor al 20%, la mayoría de estos son aminoácidos hidrofóbicos, tales como Trp, Leu, Ile y Val, los cuales comprenden 9 de los catorce aminoácidos ubicados dentro de la región central de la proteína. El resto de los aminoácidos son aminoácidos con cadena lateral polar: Asp12, Arg37, Gln41, Cys42 y Arg45. La Figura 22 permite observar la distribución de estos aminoácidos ubicados en el centro de la estructura de nuestro modelo de hR1A. La Figura 23 señala la posición de los aminoácidos ocultos hidrofóbicos y los polares.

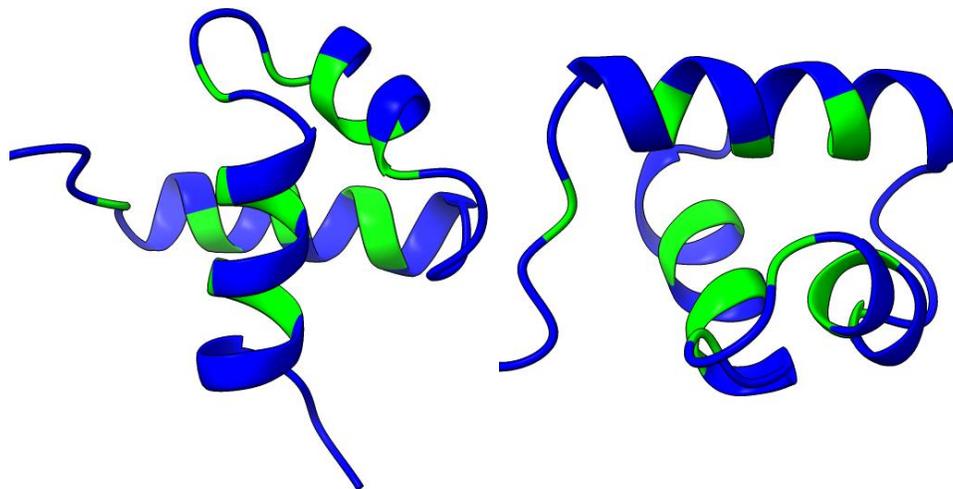


Figura 22. Representación visual de la posición de los aminoácidos enterrados (verde) y expuestos (azul).

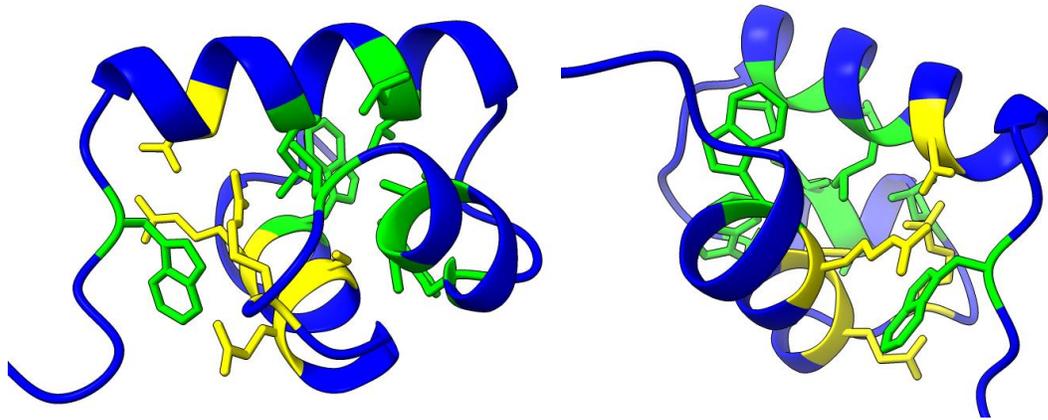


Figura 23. Representación visual de los aminoácidos ocultos hidrofóbicos (verde) y polares (amarillo). Los aminoácidos en verde constituyen el núcleo hidrofóbico de la proteína.

Podemos considerar al conjunto de aminoácidos con cadenas laterales hidrofóbicas y con un bajo RSA como el núcleo hidrofóbico de nuestra proteína.

Análisis de interacciones.

Interacciones hidrofóbicas

El servidor *PIC: Protein Interaction Calculator* indicó la presencia de interacciones hidrofóbicas que se muestran en la tabla 13. Puede observarse que el aminoácido con mayor número de interacciones es la Leu15, aminoácido con un RSA menor a 10% y es particularmente notable que este aminoácido presenta interacciones hidrofóbicas con dos de los tres triptófanos conservados de R1: Trp27 y Trp46. Otro aminoácido que comparte interacciones con estos triptófanos es Val19.

El triptófanos en la posición 7 parece no tener interacciones con otros aminoácidos dentro del núcleo hidrofóbico. Para verificar el tipo de interacciones que presenta el Trp7, se empleó nuevamente el servidor *PIC*. Los resultados se resumen en la Tabla 14.

Tabla 13: Interacciones hidrofóbicas calculadas por el servidor PIC.

Posición	Residuo	↔	Posición	Residuo
15	LEU	Interacciona con:	19	VAL
15	LEU		27	TRP
15	LEU		30	ILE
15	LEU		34	LEU
15	LEU		46	TRP
18	LEU		30	ILE
18	LEU		34	LEU
19	VAL		27	TRP
19	VAL		30	ILE
19	VAL		46	TRP
27	TRP		30	ILE
27	TRP		46	TRP
30	ILE		34	LEU

Tabla 14: Interacciones de TRP7

Residuo	↔	Residuo	Tipo de interacción	Distancia (Å)
TRP7		ARG45	Catión-Pi	4.99

Se detectó únicamente una interacción para el residuo Trp7; esta interacción se da con el residuo Arg45 ((Figura 25, c)), el cual es uno de los residuos polares con un RSA calculada menor al 20%. La naturaleza de ambos aminoácidos hace que la interacción sea de tipo catión- π . El servidor PIC considera que existe una interacción si el sistema π y el catión se encuentran a una distancia menor a 6 Å.

Interacciones entre aminoácidos: análisis con INTAA

Para este análisis se emplearon las condiciones predeterminadas en el servidor:

- Campos de fuerza: AMBER parm99
- Ambiente: Disolvente (OBC-II)
- Análisis de interacciones entre cadenas laterales (Tabla 15) y entre residuos (Tabla 16).

La Figura 24 muestra la energía de interacción neta entre las cadenas laterales de cada región de la cadena.



Figura 24. Esquema generado por el servidor INTAA, en este se observa la energía de interacción neta entre las cadenas laterales de cada región de la proteína. En azul se observan las interacciones con energía neta positiva y en rojo las interacciones con energía neta negativa.

En la Tabla 15 puede observarse que la mayoría de las interacciones entre cadenas laterales tienen una energía neta negativa, detectándose únicamente un residuo cuya cadena lateral presenta un valor no negativo, Gly23. Las energías de residuales negativas indican una interacción favorable para la estabilidad de la estructura, por lo que podemos determinar que la mayoría de las interacciones entre las cadenas laterales estabilizan al modelo. La energía residual de interacción de Gly23 es cero, por lo que es la de mayor energía de interacción calculada por el servidor. En el caso de las interacciones residuo-residuo la suma total de la energía residual de las interacciones dio un resultado negativo en todos los casos.

Los tres residuos que más aportan a la estabilidad mediante las interacciones de sus cadenas laterales son Arg45, Glu11 y Asp12. Éstos son aminoácidos con cadenas laterales polares cargadas; por lo que sus interacciones son del tipo iónica o polar ya sea con otros aminoácidos o con el disolvente.

Tabla 15: Energía residual de las interacciones cadena lateral-cadena lateral.

Residuo	Posición	Energía	Residuo	Posición	Energía
ARG	45	-190.71	HIS	44	-26.12
ASP	12	-174.92	HIS	33	-25.11
GLU	11	-154.55	VAL	5	-22.73
ARG	37	-101.92	LEU	29	-19.74
LYS	14	-78.22	GLN	21	-18.10
TRP	46	-77.46	LYS	48	-17.23
TRP	7	-73.63	GLU	53	-16.88
TRP	27	-66.51	SER	38	-16.68
GLN	43	-45.98	VAL	49	-16.65
THR	28	-45.13	THR	24	-16.43
HIS	22	-43.21	PRO	52	-14.93
GLN	47	-42.99	ARG	9	-13.58
GLN	41	-41.32	PHE	40	-13.07
LEU	15	-39.97	ASP	10	-11.51
ASP	26	-39.34	ARG	4	-11.04
ASP	13	-38.61	ASP	39	-10.29
ILE	30	-38.58	ALA	31	-6.76
LYS	16	-38.08	LYS	6	-5.25
LEU	50	-35.64	ASN	3	-3.67
ASN	51	-32.91	ASN	36	-3.02
LYS	17	-32.85	GLN	35	-2.98
LEU	34	-32.79	THR	8	-2.62
LEU	18	-31.61	SER	32	-1.92
VAL	19	-31.42	TRP	2	-1.79
GLU	20	-30.29	ASP	25	-1.34
CYS	42	-28.06	GLY	23	0.00

Por otro lado, el análisis de las interacciones residuo-residuo (Tabla 16) indica que los residuos cuyas interacciones afectan más la estabilidad de la estructura son Glu11, Asp12 y Trp46. Este último presenta interacciones de tipo hidrofóbica y de tipo catión- π . Los triptófanos conservados de Myb (Trp7, Trp27 y Trp46) se encuentran entre los que proveen mayor estabilidad a la proteína mediante sus interacciones cadenas lateral-cadena lateral con otros aminoácidos, así como mediante sus interacciones residuo-residuo. El análisis realizado previamente con el servidor PIC indica que la naturaleza de estas interacciones es principalmente de tipo hidrofóbica y catión- π .

Tabla 16: Energía residual de las interacciones residuo-residuo.

Residuo	Posición	Energía	Residuo	Posición	Energía
ASP	12	-315.59	LEU	34	-95.96
GLU	11	-286.58	GLU	20	-95.64
TRP	46	-241.23	ASP	10	-88.59
ARG	45	-240.10	LEU	50	-87.46
TRP	27	-228.90	HIS	33	-85.89
SER	38	-165.93	GLN	47	-85.29
GLN	43	-153.38	ALA	31	-75.40
LEU	15	-140.73	PHE	40	-75.24
TRP	7	-139.66	ARG	9	-74.11
THR	28	-134.56	SER	32	-72.93
CYS	42	-134.54	ASN	51	-69.89
ASP	13	-134.22	VAL	49	-67.22
LYS	14	-133.85	GLN	21	-61.64
LEU	18	-130.40	LYS	48	-55.40
LYS	16	-127.89	GLY	23	-52.56
ILE	30	-123.30	THR	24	-47.68
GLN	41	-122.79	VAL	5	-46.86
ASP	39	-121.28	GLU	53	-46.69
VAL	19	-117.99	ASN	3	-34.91
HIS	44	-115.82	LYS	6	-34.08
THR	8	-112.51	GLN	35	-31.89
ASP	26	-107.88	PRO	52	-28.99
LEU	29	-100.38	ASP	25	-26.65
ARG	37	-98.81	ASN	36	-23.57
LYS	17	-98.27	ARG	4	-10.33
HIS	22	-96.18	TRP	2	-6.69

El servidor INTAA permite también descomponer la energía neta de cada aminoácido en valores de energía de interacción relativa correspondientes a cada cadena lateral con la que interactúa, permitiendo entender la contribución individual de cada interacción.

En el caso de la Glu11, su mayor interacción se encuentra con Arg37, la cual se debe a la interacción entre las cargas de ambos residuos (Figura 25, a)). El mismo tipo de atracción iónica ocurre con Asp12, cuya energía de interacción menor se da con Arg45 (Figura 25, b)). Estas dos interacciones se dan entre residuos pertenecientes a la primera hélice (Glu11 y

Asp12) y residuos de la tercera hélice (Arg37 y Arg45). Es importante notar que Asp12, Arg37 y Arg45 son aminoácidos polares que el análisis de RSA determinó como ocultos (RSA < 20%) por lo que tendrían una baja exposición al disolvente, encontrándose cerca del núcleo hidrofóbico de la proteína.

Los tres triptófanos presentan interacciones con diferentes residuos. Trp7 presenta la mayor energía de interacción negativa con Arg45; esta interacción, previamente descrita, es de tipo catión- π (Figura 25, c)). Trp27 interactúa fuertemente con Gln43 (Figura 25, d)); la naturaleza de la interacción es también de tipo catión- π . Por su parte, Trp46 tiene interacciones importantes con Val19 y Asn51 (Figura 25, e)); la primera interacción es de naturaleza hidrofóbica y la interacción con Asn51 es de tipo puente de hidrógeno.

Como se mencionó previamente, estos triptófanos forman parte del núcleo hidrofóbico de la molécula, el cual presenta las interacciones detalladas en la tabla 13. INTAA permite calcular la energía residual para cada una de las interacciones. Los resultados se muestran en la tabla 18. También se incluye la energía de la interacción entre Trp7 y Arg45.

Tabla 18: Interacciones del núcleo hidrofóbico.

Interacción entre residuos				Tipo de interacción	
Posición	Residuo	Posición	Residuo	CL vs. CL	R vs. R
15	LEU	19	VAL	-3.66	-21.74
15	LEU	27	TRP	-3.01	-4.20
15	LEU	30	ILE	-3.05	-5.08
15	LEU	34	LEU	-3.97	-6.44
15	LEU	46	TRP	-6.29	-10.10
18	LEU	30	ILE	-6.28	-12.68
18	LEU	34	LEU	-5.45	-5.79
19	VAL	27	TRP	-4.95	-7.52
19	VAL	30	ILE	-3.43	-5.35
19	VAL	46	TRP	-13.29	-19.82
27	TRP	30	ILE	-7.37	-29.51
27	TRP	46	TRP	-5.28	-11.50
30	ILE	34	LEU	-2.75	-14.32
7	TRP	45	ARG	-41.54	-39.64

CL: Cadena lateral; R: Residuo

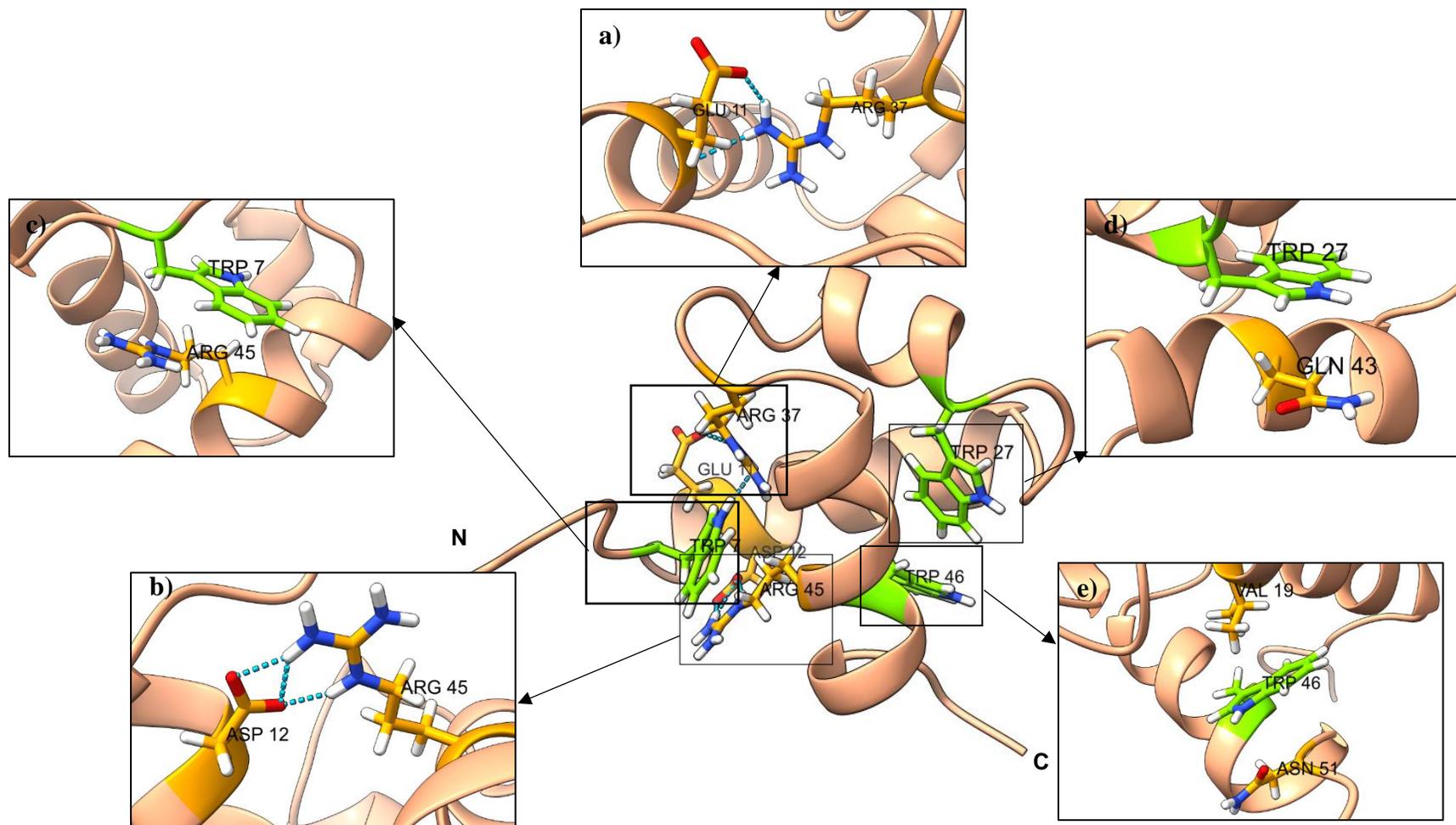


Figura 25. Visualización de las principales interacciones: a) Interacción entre Glu11 y Arg37, b) Interacción entre Asp12 y Arg45, c) Interacción del Trp7 (verde) y Arg45 (naranja), d) Interacción entre Trp27 (verde) y Gln43, e) Interacciones de Trp46 (verde) con Val19 y Asn51 (naranja). Se indican los extremos N y C terminal del fragmento.

Acoplamiento molecular.

Los resultados de las pruebas de acoplamiento molecular se describen a continuación, en todos los casos se usó como molécula objetivo la secuencia de ADN indicada en la Tabla 5

HDOCK

El servidor HDOCK generó varios modelos al identificar múltiples sitios posibles y orientaciones de interacción entre el modelo de hR1A y la molécula objetivo. La primera estructura generada fue elaborada a través de su protocolo de acoplamiento por plantillas y se muestra en la Figura 26. Este modelo empleó el PDB 6KKS como plantilla, el cual consta de una estructura con R2 y R3 unida a ADN. En él, hR1A se ha colocado en la posición que corresponde a R3 en la plantilla, observándose el posicionamiento del fragmento sobre la hendidura mayor con la tercera hélice orientada hacia la cadena de ADN.

Los 10 modelos con la mejor puntuación asignada por el servidor tienen valores muy similares y modos de unión muy diversos, algunos careciendo de la orientación determinada para las interacciones entre Myb y ADN. En la Figura 27 se muestra una superposición de los modelos generados por HDOCK. Se observa que los modelos se agrupan en dos zonas con respecto a la molécula de ADN. Estos grupos corresponden a las dos zonas donde se encuentran los sitios de unión de Myb a ADN. La posición de hR1A determinada por homología se encuentra en el grupo superior, y se observa en color blanco.

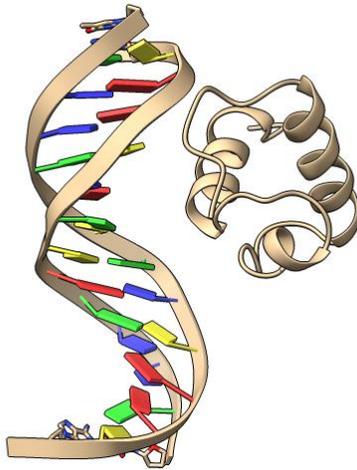


Figura 26. Modelo obtenido mediante acoplamiento molecular con plantilla

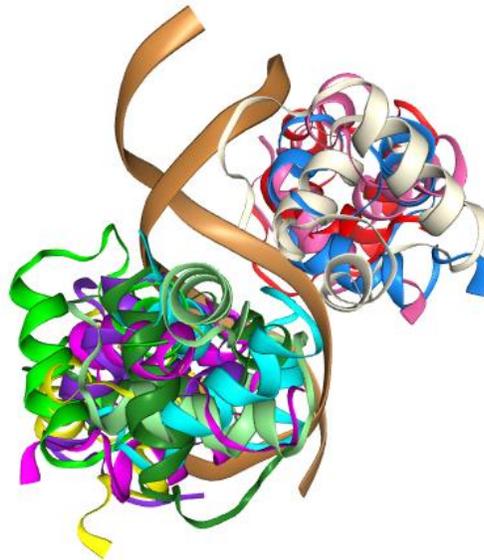


Figura 27. Superposición de los modelos generados por el servidor visualizados en HDOCK.

Analizando los modelos obtenidos, obtenemos las Figura 28 y Figura 29, donde se muestran los 6 modelos con una mayor puntuación, de acuerdo con el criterio de evaluación del servidor.

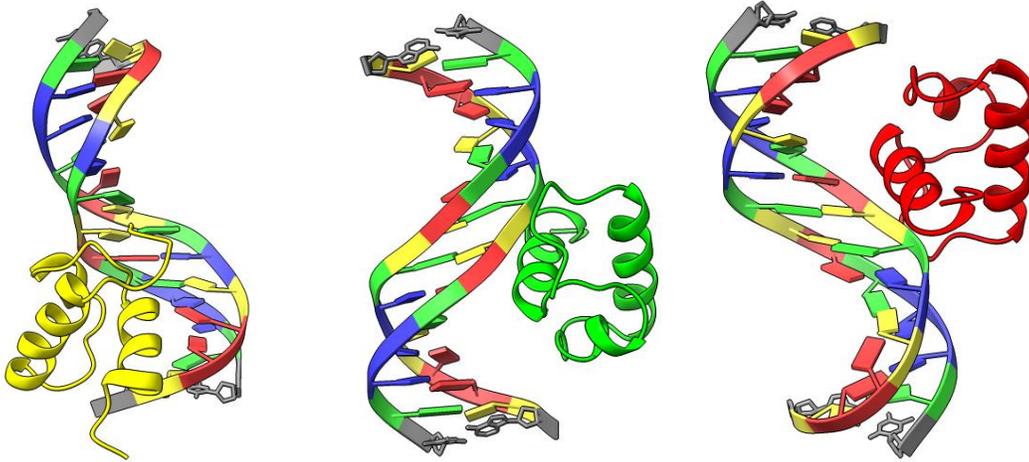


Figura 28. Modelos calificados por el servidor como aquellos con mayor puntuación.

Analizando estos tres primeros modelos:

- Modelo 1 (hR1A en amarillo): en este caso R1 se encuentra en el sitio de unión inferior, la orientación parece indicar que la tercera hélice es responsable de la interacción con el ADN.
- Modelo 2 (hR1A en verde) muestra que R1 se encuentra en una zona alejada del sitio de unión, con la hélice 3 y el segundo giro responsables de la interacción con la cadena del ADN.
- Modelo 3 (hR1A en rojo) al igual que en el modelo 2, R1 interactúa con el ADN mediante el segundo giro y la tercera hélice.

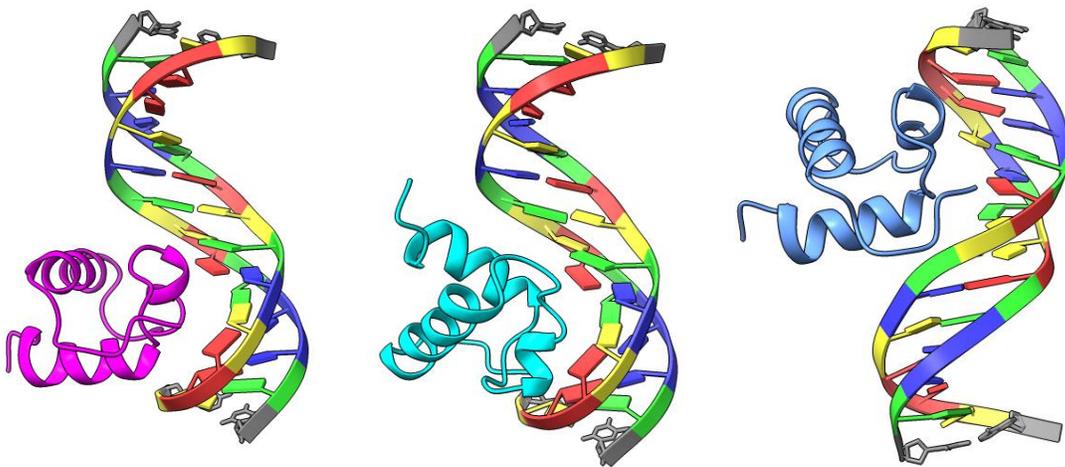


Figura 29. modelos calificados por el servidor como aquellos con mayor puntuación.

Los siguientes tres modelos con la puntuación más alta son Modelo 4 (hR1A en rosa), Modelo 5 (hR1A en cian), Modelo 6 (hR1A en azul). En estas tres estructuras la orientación de R1 sugiere que las interacciones con el ADN se darían a partir de la segunda hélice y el primer y/o segundo giro. Esto parece no coincidir con los antecedentes sobre la interacción de los fragmentos R1, R2 y R3 con el ADN, por lo cual es aconsejable ser cuidadosos al utilizar estos modelos para futuros análisis.

HADDOCK

HADDOCK genera una serie de agrupamientos y selecciona los mejores modelos de cada uno. HADDOCK clasifica estos grupos en base a la Z-score calculada, considerando como mejores a aquellos con valores más negativos. Para este estudio se tomaron los tres modelos con Z-score negativa. Estos tres modelos se muestran en la Figura 30.

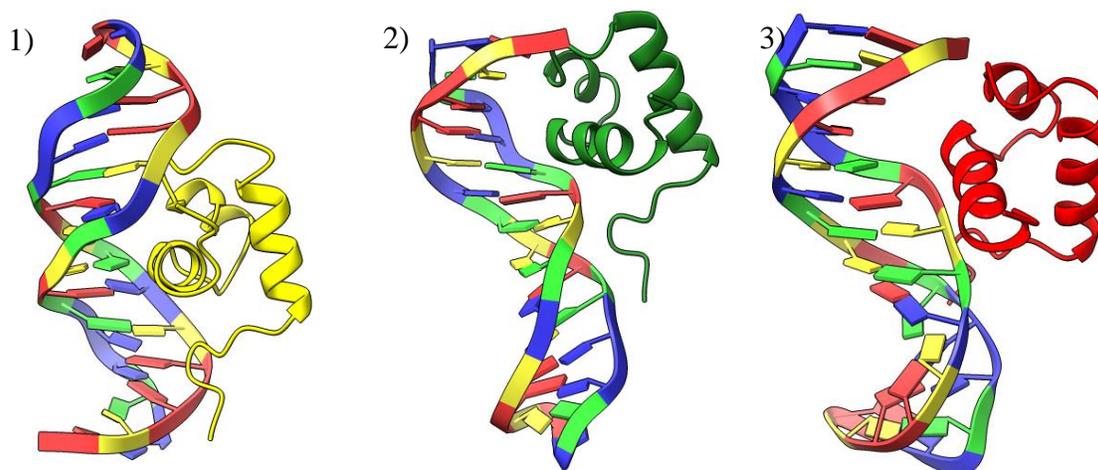


Figura 30. Tres modelos generados por HADDOCK

- 1) Grupo 2, Z-score: -1.8. Se observa que la posición de hR1A (amarillo) con respecto al ADN se encuentra en la zona intermedia entre los dos sitios de unión de Myb. R1 se encuentra en la hendidura mayor con la tercera hélice orientada hacia el ADN. Si bien la orientación de R1 es adecuada, la posición con respecto a los sitios de unión parece diferir de los antecedentes reportados.
- 2) Grupo 3, Z-score: -0.8. La posición de R1 (verde) parece encontrarse más cercana al sitio de unión que en el caso anterior.

- 3) Grupo 12, Z-score: -0.2. Al igual que el caso 2, la posición de R1 parece más adecuada al sitio de unión; sin embargo, la orientación es ligeramente diferente al caso anterior.

Es importante notar que en los tres modelos se observa una deformación considerable de la cadena de ADN empleada como objetivo, lo cual afecta la fiabilidad de los resultados obtenidos.

NPDock

Para nuestro trabajo se suministraron las mismas estructuras PDB que se suministraron a los dos anteriores servidores y se seleccionaron las opciones predeterminadas del servidor, variando únicamente el número de residuos de hR1A en interacción con la molécula blanco. Se seleccionaron los valores de 1, 5 y 10. El servidor generó los resultados observables en la Figura 31.

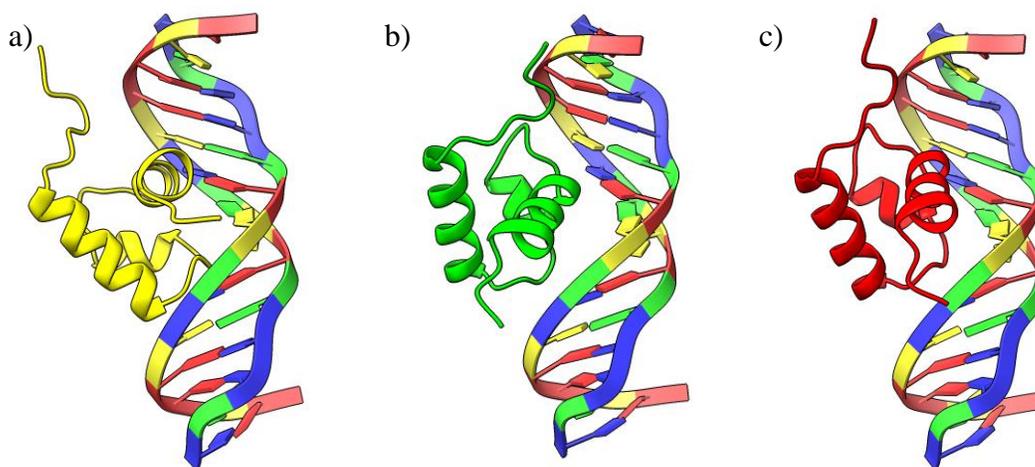


Figura 31. Estructuras con a) 1, b) 5 y c) 10 residuos en interacción con el blanco.

Observamos que a medida que el número de residuos en interacción aumenta, cambia la posición del fragmento R1. En el modelo generado con un solo residuo en interacción (hR1A en amarillo), el fragmento se encuentra situado en un punto intermedio entre los dos posibles sitios de unión. La tercera hélice y el primer giro se encuentran orientados hacia la cadena de ADN. En el modelo b), el fragmento de hR1A se sitúa de manera más cercana a la secuencia de reconocimiento, la tercera hélice continúa orientándose hacia el ADN. Finalmente, el modelo c) se encuentra en la posición más cercana al sitio de unión.

Se observa que un incremento en el número de posibles contactos entre la molécula de ADN y hR1A modifica la posición de esta última, acercándola más hacia el sitio de unión reportada para R3 sobre esa secuencia de ADN. Es importante notar que, al igual que en los resultados de HADDOCK, la molécula de ADN presenta ciertas deformaciones estructurales, siendo estas mucho más ligeras que las observadas en los resultados del caso anterior.

Los resultados de las tres pruebas de acoplamiento molecular con la estructura del repetido R1 modelada fueron diferentes en cada uno de los servidores utilizados, no generando una posición de interacción consistente. Esto puede deberse a varios factores, en primer lugar, fue posible observar que la estructura del ADN proporcionada fue deformada en las pruebas realizadas en los servidores HADDOCK y NPdock, siendo estos particularmente notables en las pruebas con HADDOCK, donde las dos cadenas de nucleótidos de ADN fueron unidas durante el proceso. Estas deformaciones afectan los cálculos de la interacción, disminuyendo la fiabilidad de los resultados obtenidos. Por otro lado, si bien no es posible observar a primera vista deformaciones en la estructura de hR1A, es posible que de igual forma hayan ocurrido cambios en su estructura durante el procesamiento por parte de los servidores, principalmente en la conformación de las cadenas laterales de los aminoácidos, donde existe mayor grado de libertad. Es también importante recordar que la secuencia “canónica” de unión al ADN empleada en las pruebas de acoplamiento molecular no es la única secuencia de reconocimiento detectada para las proteínas Myb, por lo que no es posible concluir que el repetido R1 no interactúa con el ADN usando únicamente los datos obtenidos en estas tres pruebas de acoplamiento molecular. Es necesario llevar a cabo más pruebas que actúen como control para corroborar estos resultados.

6.2. Obtención de R1

6.2.1. Transformación y expresión.

Se transformaron exitosamente cepas de *E. coli* BL21-DE3 y Rosetta con el plásmido de hR1A (Figura 7). Los resultados obtenidos en las pruebas de expresión mostraron que ambas

cepas presentaban niveles de expresión muy similares en la prueba de 37 °C (Figura 32) sin embargo, las pruebas a 30 °C mostraron que la cepa de Rosetta™ presentaba una expresión muy pobre a esta temperatura (Figura 33). En ambas figuras se indica con un recuadro rojo la zona en la cual se esperaba la presencia de la proteína, la cual cuenta con una masa molecular de 8248 g/mol (8.25 kDa)

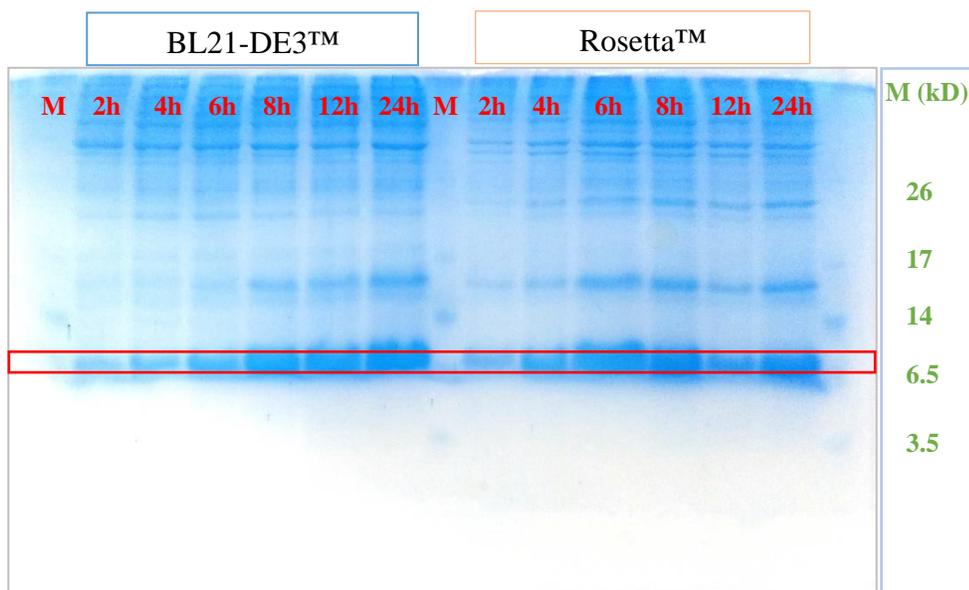


Figura 32: Gel de electroforesis de las pruebas de expresión de repetido R1 de A-Myb en bacterias a 37 °C. Se indican los carriles utilizados por cada muestra y el marcador de peso molecular (M) así como los pesos de los estándares. El recuadro rojo indica la posición en la que se esperaba la proteína expresada (MM=8.25 kDa)

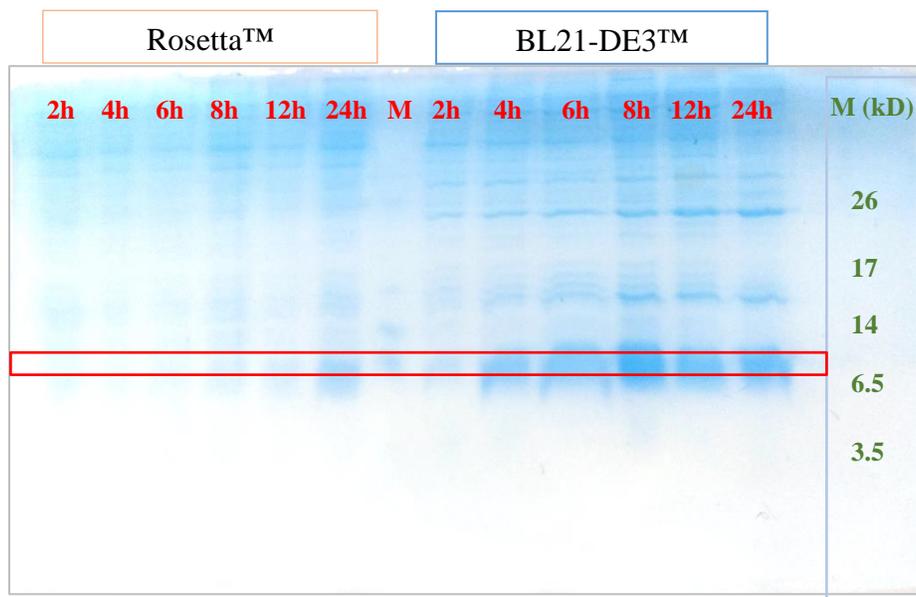


Figura 33: Gel de electroforesis de las pruebas de expresión de repetido R1 de A-Myb en bacterias a 30 °C. Se indican los carriles utilizados por cada muestra y el marcador de peso molecular (M) así como los pesos de los estándares. El recuadro rojo indica la posición en la que se esperaba la proteína expresada (MM=8.25 kDa)

La mayor producción de proteína se observó en la prueba de expresión a 37 °C. Se determinó que el tiempo óptimo de crecimiento se encuentra en las ocho horas. a partir de ese momento la intensidad de los puntos correspondientes a hR1A permaneció constante en la expresión a 37 °C, y se observó una pérdida en la intensidad en el caso de la expresión a 30 °C. También se observó un incremento en la intensidad de las bandas fuera de la zona donde se esperaría la proteína, posiblemente por la formación de péptidos o agregados proteicos no deseados.

Con estos resultados se estandarizó la metodología para la producción de hR1A con los siguientes parámetros:

- Cepa de *E. coli*: BL21-DE3.
- Temperatura de expresión: 37 °C.
- Tiempo de expresión: 8 h.

7. Conclusiones y perspectivas

Se obtuvo un modelo de la estructura terciaria del repetido R1 del dominio de unión al ADN de la proteína A-Myb de humano donde la distribución espacial de los residuos determinó la existencia de un núcleo hidrofóbico dentro de la estructura de hR1A.

Análisis de las interacciones intramoleculares mostraron que el hR1A contiene fuertes interacciones hidrofóbicas, principalmente dentro del núcleo hidrofóbico, e interacciones iónicas que la proveen de estabilidad estructural.

Los resultados de acoplamiento molecular no mostraron una posición de interacción consistente entre hR1A y el fragmento de ADN, por lo que no fue posible determinar, con los datos obtenidos, la presencia de una región en el repetido R1 que interactúe directamente con el fragmento “canónico” de interacción con el ADN.

Como perspectivas a corto plazo se plantea llevar a cabo pruebas de acoplamiento molecular variando la secuencia de ADN empleada como objetivo, así como llevando a cabo controles comparativos de acoplamientos empleando estructura reportadas de los repetidos R2 y R3 como ligantes.

Entre los objetivos a futuro, se contempla finalizar con la fase experimental para obtener el repetido R1. De ser posible su obtención, se llevarían a cabo pruebas *in vitro* de interacción con ADN.

8. Bibliografía y referencias

1. McMurry J. *Química Orgánica*. 7a Edición. CENGAGE Learning; 2008.
2. Gromiha MM. Protein Stability. In: *Protein Bioinformatics*. Elsevier; 2010:209-245. doi:10.1016/b978-8-1312-2297-3.50006-0
3. Williamson M. *How Proteins Work*. Garland Science; 2012. doi:10.1201/9781136665493
4. Mahmoud Al-Hejin A, Singh Bora R, Morsi M. Ahmed M. Plasmids for Optimizing Expression of Recombinant Proteins in E. coli . In: *Plasmid*. IntechOpen; 2019. doi:10.5772/intechopen.82205
5. Clark DP, Pazdernik NJ. *Biotechnology*. Segunda ed. Elsevier; 2016.
6. Gräslund S, Nordlund P, Weigelt J, et al. Protein production and purification. *Nat Methods*. 2008;5(2):135-146. doi:10.1038/nmeth.f.202
7. Nelson DL, Cox MM. *Lehninger Principles of Biochemistry*. W.H. Freeman & Company; 2008. doi:10.1007/978-3-662-08289-8
8. Hillenkamp F, Karas M, Beavis RC, Chait BT. Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry of Biopolymers. *Anal Chem*. 1991;63(24):1193 A-1202 A. doi:10.1021/ac00024a002
9. Karas M, Krüger R. Ion formation in MALDI: The cluster ionization mechanism. *Chem Rev*. 2003;103(2):427-439. doi:10.1021/cr010376a
10. Kelly SM, Jess TJ, Price NC. How to study proteins by circular dichroism. *Biochim Biophys Acta - Proteins Proteomics*. 2005;1751(2):119-139. doi:10.1016/j.bbapap.2005.06.005
11. Berg JM, Tymoczko JL, Stryer L. Three-Dimensional Protein Structure Can Be Determined by NMR Spectroscopy and X-Ray Crystallography. Published online 2002.
12. Calabrese B. Cloud-Based Molecular Modeling Systems. In: *Encyclopedia of Bioinformatics and Computational Biology*. Elsevier; 2019:261-264. doi:10.1016/B978-0-12-809633-8.20378-6
13. Gromiha MM, Nagarajan R, Selvaraj S. Protein Structural Bioinformatics: An Overview. In: *Encyclopedia of Bioinformatics and Computational Biology*. Elsevier; 2019:445-459. doi:10.1016/B978-0-12-809633-8.20278-1
14. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583-589. doi:10.1038/s41586-021-03819-2
15. Bonneau R, Baker D. Ab Initio Protein Structure Prediction: Progress and Prospects. *Annu Rev Biophys Biomol Struct*. 2001;30(1):173-189. doi:10.1146/annurev.biophys.30.1.173
16. Roy K, Kar S, Das RN. *Other Related Techniques.*; 2015. doi:10.1016/b978-0-12-801505-6.00010-7
17. Arratia J, Aguirre J. Los factores de transcripción tipo Myb, una familia de reguladores de la diferenciación celular conservada en los organismos eucariontes. *TIP*. 2013;16(2):98-108. doi:10.1016/S1405-888X(13)72081-3
18. Latchman DS. Transcription factors: an overview. *Int J Exp Pathol*. 1993;74(5):417-422. doi:10.1016/B0-12-370879-6/00386-0
19. Ness SA. The Myb oncoprotein: regulating a regulator. *Biochim Biophys Acta - Rev Cancer*. 1996;1288(3):F123-F139. doi:10.1016/S0304-419X(96)00027-3

20. Bergholtz S. The highly conserved DNA-binding domains of A-, B- and c-Myb differ with respect to DNA-binding, phosphorylation and redox properties. *Nucleic Acids Res.* 2001;29(17):3546-3556. doi:10.1093/nar/29.17.3546
21. George OL, Ness SA. Situational awareness: Regulation of the myb transcription factor in differentiation, The cell cycle and oncogenesis. *Cancers (Basel)*. 2014;6(4):2049-2071. doi:10.3390/cancers6042049
22. Oh I-H, Reddy EP. The myb gene family in cell growth, differentiation and apoptosis. *Oncogene*. 1999;18(19):3017-3033. doi:10.1038/sj.onc.1202839
23. Weston K. Myb proteins in life, death and differentiation. *Curr Opin Genet Dev*. 1998;8(1):76-81. doi:10.1016/S0959-437X(98)80065-8
24. Fuglerud BM, Lemma RB, Wanichawan P, Sundaram AYM, Eskeland R, Gabrielsen OS. A c-Myb mutant causes deregulated differentiation due to impaired histone binding and abrogated pioneer factor function. *Nucleic Acids Res.* 2017;45(13):7681-7696. doi:10.1093/nar/gkx364
25. Mitra P. Transcription regulation of MYB: a potential and novel therapeutic target in cancer. *Ann Transl Med.* 2018;6(22):443-443. doi:10.21037/atm.2018.09.62
26. Liu X, Xu Y, Han L, Yi Y. Reassessing the potential of MYB-targeted anti-cancer therapy. *J Cancer*. 2018;9(7):1259-1266. doi:10.7150/jca.23992
27. Ogata K, Morikawa S, Nakamura H, et al. Comparison of the free and DNA-complexed forms of the DMA-binding domain from c-Myb. *Nat Struct Mol Biol.* 1995;2(4):309-320. doi:10.1038/nsb0495-309
28. Tanikawa J, Yasukawa T, Enari M, et al. Recognition of specific DNA sequences by the c-myb protooncogene product: role of three repeat units in the DNA-binding domain. *Proc Natl Acad Sci.* 1993;90(20):9320-9324. doi:10.1073/pnas.90.20.9320
29. Saikumar P, Murali R, Reddy EP. Role of tryptophan repeats and flanking amino acids in Myb-DNA interactions. *Proc Natl Acad Sci.* 1990;87(21):8452-8456. doi:10.1073/pnas.87.21.8452
30. Biedenkapp H, Borgmeyer U, Sippel AE, Klempnauer KH. Viral myb oncogene encodes a sequence-specific DNA-binding activity. *Nature*. 1988;335(6193):835-837. doi:10.1038/335835a0
31. Howe KM, Watson RJ. Nucleotide preferences in sequence-specific recognition of DNA by c-myb protein. *Nucleic Acids Res.* 1991;19(14):3913-3919. doi:10.1093/nar/19.14.3913
32. Weston K. Extension of the DNA binding consensus of the chicken c-Myb and v-Myb proteins. *Nucleic Acids Res.* 1992;20(12):3043-3049. doi:10.1093/nar/20.12.3043
33. Drozdetskiy A, Cole C, Procter J, Barton GJ. JPred4: A protein secondary structure prediction server. *Nucleic Acids Res.* 2015;43(W1):W389-W394. doi:10.1093/nar/gkv332
34. Yachdav G, Kloppmann E, Kajan L, et al. PredictProtein - An open resource for online prediction of protein structural and functional features. *Nucleic Acids Res.* 2014;42(W1):337-343. doi:10.1093/nar/gku366
35. Källberg M, Margaryan G, Wang S, Ma J, Xu J. Raptorx server: A resource for template-based protein structure modeling. *Methods Mol Biol.* 2014;1137:17-27. doi:10.1007/978-1-4939-0366-5_2
36. Bernhofer M, Dallago C, Karl T, et al. PredictProtein - Predicting Protein Structure and Function for 29 Years. *Nucleic Acids Res.* Published online 2021:1-5.

- doi:10.1093/nar/gkab354
37. Waterhouse A, Bertoni M, Bienert S, et al. SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* 2018;46(W1):W296-W303. doi:10.1093/nar/gky427
 38. Haddad Y, Adam V, Heger Z. Ten quick tips for homology modeling of high-resolution protein 3D structures. *PLoS Comput Biol.* 2020;16(4):1-19. doi:10.1371/journal.pcbi.1007449
 39. Yang J, Zhang Y. Protein Structure and Function Prediction Using I-TASSER. *Curr Protoc Bioinforma.* 2015;52(1):96-108. doi:10.1002/0471250953.bi0508s52
 40. Krieger E, Joo K, Lee J, et al. Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. *Proteins Struct Funct Bioinforma.* 2009;77(SUPPL. 9):114-122. doi:10.1002/prot.22570
 41. Chen VB, Arendall WB, Headd JJ, et al. MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallogr Sect D Biol Crystallogr.* 2010;66(1):12-21. doi:10.1107/S0907444909042073
 42. Ali S, Hassan M, Islam A, Ahmad F. A Review of Methods Available to Estimate Solvent-Accessible Surface Areas of Soluble Proteins in the Folded and Unfolded States. *Curr Protein Pept Sci.* 2014;15(5):456-476. doi:10.2174/1389203715666140327114232
 43. Swindells MB. A procedure for the automatic determination of hydrophobic cores in protein structures. *Protein Sci.* 1995;4(1):93-102. doi:10.1002/pro.5560040112
 44. Tina KG, Bhadra R, Srinivasan N. PIC: Protein Interactions Calculator. *Nucleic Acids Res.* 2007;35(SUPPL.2):473-476. doi:10.1093/nar/gkm423
 45. Galgonek J, Vymětal J, Jakubec D, Vondrášek J. Amino Acid Interaction (INTAA) web server. *Nucleic Acids Res.* 2017;45(W1):W388-W392. doi:10.1093/nar/gkx352
 46. Yan Y, Zhang D, Zhou P, Li B, Huang SY. HDock: A web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy. *Nucleic Acids Res.* 2017;45(W1):W365-W373. doi:10.1093/nar/gkx407
 47. De Vries SJ, Van Dijk M, Bonvin AMJJ. The HADDOCK web server for data-driven biomolecular docking. *Nat Protoc.* 2010;5(5):883-897. doi:10.1038/nprot.2010.32
 48. Tuszynska I, Magnus M, Jonak K, Dawson W, Bujnicki JM. NPDock: A web server for protein-nucleic acid docking. *Nucleic Acids Res.* 2015;43(W1):W425-W430. doi:10.1093/nar/gkv493
 49. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol.* 1982;157(1):105-132. doi:10.1016/0022-2836(82)90515-0