UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias Biológicas y Ambientales

Machine learning y biología computacional aplicados al reconocimiento de sitios de regulación en la ruta del ácido jasmónico-isoleucina en *Arabidopsis thaliana*

Proyecto de investigación

Samara Mishelle Oña Chuquimarca

Ingeniería en Procesos Biotecnológicos

Trabajo de titulación presentado como requisito
para la obtención del título de
Ingeniera en Procesos Biotecnológicos

Quito, 25 de abril de 2018

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

COLEGIO DE CIENCIAS BIOLÓGICAS Y AMBIENTALES

HOJA DE CALIFICACIÓN
DE TRABAJO DE TITULACIÓN

Machine learning y biología computacional aplicados al reconocimiento de sitios de regulación en la ruta del ácido jasmónico-isoleucina en *Arabidopsis thaliana*.

Samara Mishelle Oña Chuquimarca

Calificación:

Nombre del profesor, Título académico Miguel Ángel Méndez, Ph.D.

Firma del profesor _____

Quito, 25 de abril de 2018

Derechos de Autor

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior.

Firma del estudiante: _____

Nombres y apellidos: Samara Mishelle Oña Chuquimarca

Código: 00116026

Cédula de Identidad: 1721152344

Lugar y fecha: Quito, 25 de abril de 2018

RESUMEN

El ácido jasmónico (JA) es un compuesto orgánico que desempeña un papel importante en la inmunidad de las plantas. Cuando la concentración de JA aumenta, se activa la expresión génica de las vías moleculares de defensa debido al desacoplamiento de las proteínas represoras JAZ y el factor de transcripción bHLH MYC. El objetivo principal de este estudio es identificar los residuos más importantes en términos energéticos (*hotspots*) responsables de la interacción JAZ-MYC. Se crearon modelos 3D validados computacionalmente de los doce complejos JAZ-MYC3 mediante modelación por homología. Estos modelos fueron procesados en una rutina de dinámica molecular. Luego se procedió a analizar los resultados para obtener dos tipos de predicciones: la energía libre de unión promedio de los doce complejos que fue de -10.94 +/-2.67 kJ/mol; y los *hotspots* moleculares, obtenidos vía técnicas de machine learning, que fueron en su mayoría aminoácidos cargados o aromáticos localizados en la interface proteína-proteína. Finalmente, se destaca el descubrimiento del motivo conservado SL••FL•••R como un posible sitio de reconocimiento molecular compartido en todas las proteínas JAZs.

Palabras clave: JAZ, MYC, defensa plantas, machine learning, biología computacional, dinámica molecular, *hotspots*.

ABSTRACT

Jasmonic acid (JA) is a volatile organic compound that plays an essential role in the immunity of plants. When the concentration of JA increases, the gene expression of the molecular defense pathway is activated due to the decoupling of the repressor proteins JAZ and the transcription factor bHLH MYC. The primary objective of this study is to identify the most critical residues regarding energy (hotspots) responsible for JAZ-MYC interaction. We solved and validated twelve JAZ-MYC3 3D *in silico* structures by homology modeling. These models were introduced in a molecular dynamics pipeline to obtain two predictions: average binding free energy was -10.94 +/-2.67 kJ/mol for the twelve complexes; and the molecular hotspots, predicted by machine learning techniques, which in most extent were charged or aromatic amino acids that are located in the protein-protein interface. Finally, we highlight the discovery of the conserved SL••FL•••R motif as a possible shared molecular recognition site in all JAZs proteins.

Key words: JAZ, MYC, plant defense, machine learning, computational biology, molecular dynamics, *hotspots.*

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

El sistema inmune de las plantas

Las plantas son organismos de vida sésil que a diferencia de los animales no pueden emplear el movimiento como táctica defensiva ante estímulos físicos y químicos nocivos del ambiente. Por lo tanto, a lo largo de su evolución han desarrollado un sistema de señalización molecular complejo y especializado que les permite responder de manera efectiva al ataque de herbívoros, parásitos vegetales, microorganismos patógenos y compuestos perjudiciales abundantes en el medio ambiente.

El jasmonato (JA), etileno y ácido salicílico son las principales fitohormonas de señalización molecular dentro del metabolismo vegetal (Memelink, 2009). Cada una de ellas estimula o inhibe cascadas de señalización clave para el funcionamiento de la planta (Memelink, 2009). El ácido jasmónico es el responsable de inducir la respuesta en contra de ácaros, hongos necrotróficos e insectos. El ácido salicílico es una importante defensa contra bacterias y organismos biotróficos. El etileno actúa regulando las rutas del ácido jasmónico y ácido salicílico. En este estudio nos enfocaremos en la ruta del jasmonato y su papel en la activación del sistema de defensa de las plantas.

La ruta de señalización del jasmonato es responsable de coordinar y dirigir la formación de complejos proteicos que promueven o reprimen la transcripción de genes involucrados en la activación del sistema inmune de la planta como por ejemplo la síntesis de antocianinas (Katsir et al., 2008). En específico, este trabajo estudiará la interacción entre las doce proteínas represoras Jasmonate-ZIM Domain (JAZ) y el factor de transcripción basic helix-loop-helix (bHLH) MYC que responden a la señalización por jasmonato.

Las proteínas JAZ son un conjunto de proteínas homólogas, redundantes y desordenadas reguladas por la fitohormona JA (Staswick, 2008). MYC es un factor de transcripción conservado con un dominio bHLH que interactúa con el motivo CACGTG de

G-Box localizado en los promotores de genes de activación de la respuesta inmune (Wasternack & Hause, 2013). Los complejos JAZ-MYC actúan como *switches* moleculares que permiten prender o apagar la transcripción de genes como DRF, TAT3, LOX3, GGP1, SUR1, PDF1.2, VSP2, JAZ1, JAZ3, JAZ10, JAZ7 (Schweizer et al., 2013; Pauwels & Goossens, 2011). Todos involucrados en distintas tareas de activación de mecanismos de defensa.

El funcionamiento de estos *switches* moleculares es regulado por la diferencia de concentración del conjugado de isoleucina y ácido jasmónico (JA-Ile) (Santner et al., 2007). El JA-Ile pertenece a la familia de oxilipinas, moléculas de señalización de naturaleza lipídica, que regulan el desarrollo de la planta (Santner et al., 2007). Jasmonoyl-isoleucine conjugate synthase es responsable de la síntesis de JA-Ile (Santner et al., 2007).

En ausencia de estrés biótico o abiótico, la biosíntesis de JA-Ile es baja lo cual provoca que las proteínas de represión JAZ permanezcan unidas al factor de transcripción MYC (Wasternack & Hause, 2013). El complejo JAZ-MYC recluta a los correpresores Novel Interactor of JAZ (NINJA) y TOPLESS (TPL) que incorporan a histone deacetylase 6 and 19 (HDA6 and HDA19) para reprimir la transcripción de los genes que activan la respuesta inmune (Wasternack & Hause, 2013).

Ante estímulos ambientales, la biosíntesis de JA-Ile se potencializa y la concentración de la fitohormona sube (Santner et al., 2007). Cuando esto ocurre, las proteínas JAZ y los correpresores se desacoplan del factor de transcripción MYC (Wasternack & Hause, 2013). Se forma el complejo JAZ-JA-Ile que se dirige hacia SCF (COI1) ubiquitina ligasa la cual marca a las proteínas JAZ (ubiquitinación) para que sean degradadas por el proteosoma 26S (Vasyukova & Ozeretskovskaya, 2009). Mientras tanto, MYC recluta el complejo Mediador subunidad 25 (MED25), y el resto de maquinaria de transcripción que inicia la expresión de los genes antes suprimidos (Wasternack & Hause, 2013).

Complejos proteicos JAZ-MYC

Entender la interacción proteína-proteína de los doce complejos JAZ-MYC3 es relevante para poder controlar la expresión o represión de genes involucrados en la ruta del jasmonato y de esta manera regular la respuesta de las plantas ante estímulos ambientales. Debido a que la estructura primaria es la única información disponible para la mayoría de los complejos JAZ-MYC, es necesario implementar una metodología de modelado estructural para dilucidar la estructura secundaria y terciaria de estos complejos.

La modelación estructural de complejos proteicos ha permitido acelerar el avance en el área de diseño de fármacos e ingeniería de proteínas. De hecho, las estructuras de los receptores acoplados a las proteínas G, principal diana de al menos un tercio de los fármacos aprobados por la FDA, fueron predichos computacionalmente debido a la dificultad de resolver la estructura cristalográfica de los mismos (Schmidt et al., 2014). De esta manera el primer paso del flujo de trabajo computacional es la predicción de las estructuras 3D de los doce complejos proteicos JAZ-MYC3.

La validación de los modelos estructurales es crucial para obtener resultados confiables a posteriori. Por lo tanto, luego de la obtención de los modelos JAZ-MYC3 se requiere una fase de refinación y validación de los mismos mediante herramientas de modelado molecular que evalúan las interacciones atómicas, entre residuos y la calidad general de los modelos proteicos. Los modelos estructurales finales que hayan pasado le filtro de validación serán el punto de partida para estudios de interacción más complejos.

El estudio de interacción proteína-proteína se realiza de forma experimental y computacional. Estudios experimentales se basan esencialmente en sistemas de doble híbrido que miden la interacción física proteína-proteína en cultivos de levaduras modificados genéticamente (Causier & Davies, 2002). La técnica molecular se basa en la transcripción de

un gen reportero que depende de la reconstrucción de un factor de transcripción dividido en dos dominios (Causier & Davies, 2002). El dominio de unión al ADN (BD) se une químicamente a una de las proteínas de interés, mientras que, el dominio de activación (AD) se una a otra proteína de estudio o a una biblioteca de proteínas (Causier & Davies, 2002). La unión indirecta de BD y AD activan la expresión del gen reportero que trabaja como un marcador colorimétrico de interacción proteína-proteína (Causier & Davies, 2002).

Por otro lado, estudios computacionales de interacción proteica se basan en la predicción y simulación de los mecanismos de reconocimiento molecular dirigidos por fuerzas intermoleculares e interacciones bioquímicas específicas descritos a continuación. La mecánica de interacción de dos proteínas inicia con el cambio de polaridad y el empaquetamiento de los residuos que formarán la interface (Moreira et al., 2007). Este cambio aumenta la cantidad de interacciones de van der Waals y consecuentemente la liberación de energía libre (Moreira et al., 2007). El acoplamiento de las proteínas en un solo complejo ocasiona el desplazamiento de moléculas de agua de la interface hidrofóbica y el aumento de la entropía (Moreira et al., 2007). Finalmente, se llega a un estado de equilibrio de las fuerzas de interacción electrostáticas, los puentes de hidrógeno, interacciones de van der Waals que aseguran la estabilidad del complejo por un periodo de tiempo determinado (Moreira et al., 2007).

Hotspots moleculares en interfaz JAZ-MYC

El estudio computacional de la dinámica de interacción de los complejos JAZ-MYC3 busca identificar cuáles son los aminoácidos clave o *hotspots* definidos como los residuos de mayor contribución energética dentro de la interacción proteína-proteína de los doce complejos (Bogan & Thorn, 1998). El escaneo mutacional de alanina es el gold estándar para el descubrimiento de *hotspots* en interfaces proteína-proteína (Kenneth Morrow et al., 2012).

Por lo cual utilizaremos su homólogo computacional para la predicción mediante un modelo robusto de machine learning de *hotspots* en cada una de las interfaces JAZ-MYC.

La recopilación de las características físico-químicas de complejos proteicos genera grandes cantidades de datos que a simple vista parecen no tener patrones específicos ni utilidad para determinar los *hotspots* de un complejo. Sin embargo, el valor de esta información puede ser extraído mediante técnicas de machine learning y análisis estadísticos de los datos. Machine learning es una rama de la inteligencia artificial que utiliza métodos de análisis como razonamiento Bayesiano, métodos de máxima entropía, procesos Gaussianos y máquinas de vectores de soporte para buscar patrones complejos y relevantes dentro de grandes cantidades de datos (Frank et al., 2004).

En este trabajo de titulación aprovechamos las herramientas de machine learning para reconocer los *hotspots* presentes en los doce complejos JAZ-MYC del organismo modelo *Arabidopsis thaliana* para posteriormente escalar a modelos de plantas comerciales de interés nacional. En la actualidad el ácido jasmónico es un compuesto agrícola con alto valor comercial. Por lo tanto, la idea del proyecto macro es crear de la forma más eficiente un modelo análogo que genere las mismas funciones del ácido jasmónico como regulador del sistema inmune de la planta, a menor costo de producción. Este trabajo de titulación busca generar el punto de partida para el diseño robusto del fitoregulador mediante una descripción extensiva de las interfaces proteicas que forman parte del mecanismo de regulación del jasmonato.

Objetivos

Objetivo general del trabajo de titulación

- Establecer cuáles son los aminoácidos que constituyen el sitio de reconocimiento molecular de los doce complejos JAZ-MYC y los *hotspots* conservados en todos los complejos. Importantes reguladores moleculares de la ruta del ácido jasmónico.

Objetivos específicos del trabajo de titulación

- Desarrollar y validar modelos estructurales de los 12 complejos proteicos JAZ-MYC3.
- Estandarizar una metodología para el descubrimiento de sitios maestros de interacción molecular utilizando herramientas de biología computacional.

Área de estudio

El proyecto macro se centra en el área de fitopatología y biología molecular de plantas. Una de las principales interrogantes en este campo es como mejorar la respuesta inmune de las plantas ante distintos patógenos perjudiciales para los cultivos y así evitar el uso de compuestos tóxicos como insecticidas que pueden ocasionar daños irreversibles al ambiente y a las relaciones simbióticas planta-artrópodo.

El uso de potenciadores de la respuesta inmune permite estimular a la planta para la síntesis moderada de sustancias de defensa de origen natural. Estas sustancias aparecen como respuesta al ataque de artrópodos u hongos. Sin embargo, se busca que mediante el uso de potenciadores de la respuesta inmune las plantas generen estos compuestos antes del ataque. Así, el daño por patógenos sería mínimo.

El diseño de una sustancia fitoreguladora requiere el estudio extensivo de las rutas metabólicas que utilizan las plantas para prender y apagar genes de la respuesta inmune. Los estudios computacionales son una de las mejores alternativas para las primeras fases de estudio de las interacciones proteicas y el diseño de la sustancia reguladora. Esto se debe a que los flujos de trabajo *in silico* reducen la cantidad de recursos invertidos en investigación y dan excelentes pautas para el diseño experimental posterior.

De esta manera, el presente estudio de investigación se dirige al diseño y ejecución de un flujo de trabajo *in silico* para comprender las interacciones moleculares que regulan la respuesta inmune de las plantas utilizando como modelo de estudio *Arabidopsis thaliana*. La información que se obtendrá de este trabajo de titulación será el punto de partida para el diseño de la sustancia reguladora y la metodología implementada podrá ser escalada a otras plantas de interés comercial.

Justificación

En el campo de la agricultura es común utilizar el jasmonato para estimular el sistema de defensa de las plantas y así evitar que las plagas destruyan la producción de cultivos. No obstante, el costo de 100mg de ácido jasmónico (precursor de jasmonato) es de alrededor de 100 dólares lo cual es un precio bastante elevado considerando el presupuesto del agricultor ecuatoriano (Sigma Aldrich, 2017). Este estudio pretende a largo plazo dar la pauta inicial para crear un compuesto elicitor con la misma potencia pero menor costo de producción que sean más asequibles para el tratamiento de cultivos locales.

Para cumplir con este objetivo se debe entender a nivel molecular las cascadas de interacción presentes en los mecanismos inmunológicos de las plantas. Las estructuras cristalográficas de los complejos proteicos relacionados a estos mecanismos son el punto de partida para los estudios de interacción. Al momento Protein DataBank solo posee el modelo estructural de JAZ1-MYC3, JAZ9-MYC3 y JAZ10-MYC3 (Zhang et al., 2015). De esta manera, otro de los justificativos para este estudio es la creación de modelos estructurales válidos en base a homología que puedan ser una buena aproximación a modelos cristalográficos. Los mismos que serán el punto de partida para estudios de interacción más complejos.

Se busca desarrollar y validar una metodología para el descubrimiento de *hotspots* en los complejos proteína-proteína usando machine learning. Este nivel profundo de caracterización molecular es indispensable para el diseño de candidatos que actúen como fitoreguladores de la respuesta inmune y en general cualquier tipo de sustancia activa de interés comercial (Schmidt et al., 2014). Por lo tanto, la automatización de esta metodología se justifica por su gran utilidad para el entendimiento a nivel molecular de las interacciones proteína-proteína y su papel en el descubrimiento de nuevos fármacos.

MATERIALES Y MÉTODOS

Modelación de los complejos JAZ-MYC

La construcción de los modelos individuales de las proteínas se realizó utilizando Modeller en el caso de MYC3 y Expasy en el caso de las proteínas JAZ. Modeller 9.19 y Expasy son herramientas de modelación estructural que calculan en base a homología, modelos proteicos en tres dimensiones (Webb & Sali, 2014; Bienert et al., 2016). Los modelos estructurales individuales fueron ensamblados utilizando como plantilla complejos JAZ-MYC de estructura cristalográfica conocida i.e. JAZ1-MYC3, JAZ9-MYC3, JAZ10-MYC3.

El ensamblaje consistió en un alineamiento estructural de la proteína MYC3 y cada una de las doce proteínas JAZ con sus homólogos de la estructura de referencia. El alineamiento estructural se realizó en Pymol 1.7.4.5 mediante la función ¨align¨. El proceso comienza con un alineamiento de secuencia seguido de una superposición estructural de los modelos y la estructura referencia. El alineamiento termina con ciclos de control que refinan el modelo final.

Validación de los modelos JAZ-MYC

Una vez construidos los modelos de los doce complejos se realizó un procesamiento de validación antes de iniciar los cálculos de dinámica molecular. Esto se realiza para tener una estructura suficientemente estable y con la menor cantidad de errores estructurales que afecten la dinámica. Para este proceso de validación se utilizaron cuatro servidores de validación en línea usados en estudios anteriores para validar modelos proteína-proteína (Bhutani et al., 2015).

Errat

Método de validación que utiliza una función cuadrática de error la cual evalúa las interacciones atómicas pares no covalentes dentro de una estructura proteica (Colovos & Yeates, 1993). Errat calcula y compara estadísticamente la función de error de la proteína candidata contra las funciones de error de un set de datos de 96 estructuras proteicas validadas (Colovos & Yeates, 1993). Las interacciones atómicas entre carbono (C), nitrógeno (N) y oxígeno (O) que se toman en cuenta deben tener una distancia menor a 3.5 Å (Colovos & Yeates, 1993).

ProQ

Método de validación basado en machine learning que utiliza redes neurales para predecir la calidad de un modelo proteico (Wallner & Elofsson, 2003). Para la predicción utiliza una combinación de características estructurales de las cuales se destacan: superficie accesible al solvente, contactos residuo-residuo y contactos átomo-átomo (Wallner & Elofsson, 2003). ProQ calcula dos puntuaciones, LGscore y MaxSub ambos realizan una comparación estructural entre el modelo a predecir y un set de modelos correctos (Wallner & Elofsson, 2003).

QMean4

Método de validación estructural que mide la calidad absoluta de los modelos proteicos y estima el grado de "natividad" estructural del modelo (Benkert et al., 2010). Es decir, compara las características del mismo y determina la similitud entre la estructura y un set de modelos estructurales experimentales (Benkert et al., 2010). QMean4 es una función de puntuación normalizada independiente del tamaño de la proteína por lo que puede ser usada para monómeros o complejos poliméricos (Benkert et al., 2010).

SolvX

Método de validación estructural que calcula el perfil de solvatación del complejo proteico el cual se relaciona con la accesibilidad del solvente para cada residuo dentro de una proteína

(Holm & Sander1992). Las estructuras reales poseen una solvatación menor a cero pues los complejos funcionales poseen una estructura compacta (Holm & Sander1992). El método utiliza solamente este parámetro para discriminar entre proteínas con plegamiento correcto e incorrecto (Holm & Sander1992).

Dinámica molecular de los complejos JAZ-MYC

Luego de la validación de los modelos se realizaron estudios de dinámica molecular con los doce complejos por separado. Para éste propósito se usó el programa Gromacs 5.1.4 con el campo de fuerza AMBER-03. El sistema de simulación inicial consistió en una caja cúbica solvatada con moléculas de agua SPC216, condiciones de límite periódicas, modelo de agua TIP3P, iones de Na+ y Cl- para neutralizar la carga, PME para las interacciones electrostáticas y concentración de 0.1M NaCl para simular condiciones fisiológicas (Bhutani et al., 2015).

El primer paso de minimización de energía se realizó con el algoritmo steepest descent con 10 kJ/mol/nm como máxima fuerza de convergencia del sistema. El segundo paso fue una dinámica de acoplamiento de temperatura (NVT) utilizando el termostato V-rescale para subir gradualmente la temperatura hasta 310K. La dinámica NVT se realizó utilizando restricciones de movimiento de la proteína para facilitar la relajación de las moléculas de agua en todo el sistema y un tiempo de simulación de 500ps. El tercer paso fue una dinámica de acoplamiento de presión (NPT) en la que se quitaron las restricciones y se utilizó el baróstato Parrinello-Rahman para estabilizar la presión del sistema a 1 bar. Se usó un valor de compresibilidad de $4.5x10^{-5}$ bar^{-1} y un tiempo de simulación de 500ps (Bhutani et al., 2015).

En última instancia se realizó la dinámica de producción que es el cálculo principal en el estudio de dinámica molecular y genera un archivo de trayectoria que se usa para el análisis a fondo del sistema. La dinámica de producción se realizó con las mismas

condiciones de presión y temperatura (1 bar, 310K) por un tiempo de simulación de 50ns. Se analizó la convergencia del sistema a lo largo de la dinámica y características generales de calidad con las herramientas de análisis de trayectoria integradas en Gromacs. A saber, temperatura, presión, energía, volumen, densidad, dimensiones de la caja de simulación, RMSF (root mean square fluctuation), RMSD (root mean square deviation), radio de giro, número de puentes de hidrógeno (Bhutani et al., 2015).

Energía libre de unión de los complejos JAZ-MYC

La predicción se realizó utilizando el programa FoldX Suite, funcionalidad AnalyseComplex, que permite el cálculo de la energía libre de interacción entre dos complejos proteicos ($\Delta G_{Unión}$). El cálculo de la energía de Gibbs se realiza a cada proteína por separado y al complejo (Schymkowitz et al., 2005). Se calcula la energía libre mediante la siguiente fórmula:

$$\Delta G_{\mathrm{binding}} = \Delta G_{\mathrm{AB}} - (\Delta G_A + \Delta G_B)$$

Donde $\Delta G_{\mathrm{binding}}$ representa el cambio de energía libre de unión total del complejo; ΔG_{AB}, el cambio de energía libre de Gibbs del complejo; ΔG_A, el cambio de energía libre de Gibbs de la proteína A y ΔG_B el cambio de energía libre de Gibbs de la proteína B (Schymkowitz et al., 2005).

Hotspots moleculares de los complejos JAZ-MYC

La predicción de *hotspots* moleculares se realizó mediante técnicas de machine learning. Para ello se recolectó un set de 20 datos experimentales sobre mutaciones puntuales dirigidas realizadas en los complejos JAZ1-MYC3 y JAZ9MYC3 (**Tabla 2**). Este set de entrenamiento se utilizó para alimentar a los algoritmos de predicción y escoger el más preciso para descubrir *hotspots* dentro del sistema de estudio.

Los cálculos energéticos que se utilizaron como descriptores de cada una de las instancias del set de entrenamiento y del set de prueba fueron realizados con el servidor en

línea FoldX Suite en específico la funcionalidad PSSM (Schymkowitz et al., 2005). Esta herramienta permite determinar el $\Delta\Delta G_{unión}$ luego de una mutación puntual (Schymkowitz et al., 2005). Se realizó mutaciones para los 20 aminoácidos esenciales para todas las posiciones que comprenden la interface JAZ-MYC de los 12 complejos.

El flujo de trabajo iterativo para probar los distintos algoritmos de machine learning fue realizado en Waikato Environment for Knowledge Analysis (WEKA) v.3.8.1. Se utilizaron los algoritmos Random Forest, Multilayer Perceptron, Naive Bayes (NB) y Sequential minimal optimization (SMO) lo cuáles han sido probados con éxito en la predicción de *hotspots* (Aguilera-Pesantes et al., 2017; Liu et al., 2009; Wang et al., 2012; Murakami & Mizuguchi, 2010; Mintseris, 2003). Para medir el desempeño de estos algoritmos se utilizaron métricas descriptivas tales como exactitud, precisión, exhaustividad y puntaje F1. Además, se realizó una reducción de dimensionalidad de los datos para entrenar al modelo con la cantidad mínima de descriptores con mayor poder predictivo.

Resultados

Modelos 3D validados de los complejos JAZ-MYC

Los modelos estructurales obtenidos de las herramientas Modeller y Expasy que posteriormente fueron optimizados mediante dinámica molecular se encuentran representados en la **Figura 1**. El modelo de la proteína MYC3 contiene los aminoácidos 44-236 que corresponden al dominio JID-TAD el cual tiene forma globular y está formada por una estructura de hojas plegadas beta rodeada por 5 hélices alfa. Los modelos de las proteína JAZ 1-12 corresponden al dominio Jas que contiene, dependiendo de la proteína, entre 15-20 aminoácidos ordenados en una hélice alfa. En todos los doce complejos, las proteínas JAZ se encuentran formando interface con dos de las hélices de MYC3 que forman una cavidad en forma de medialuna.

La validación cuantitativa de los modelos estructurales se encuentra en las **Figuras 2-7**. Ésta se llevó a cabo en tres etapas de la dinámica molecular. La primera validación (V1) fue del modelo inicial obtenido inmediatamente después del alineamiento estructural y la optimización de las zonas lazo (loop) del complejo. Se evaluó este modelo para tener una línea base que permita observar el progreso de la optimización estructural del modelo. La segunda validación (V2) se realizó después del cálculo de energía de minimización en el cuál la estructura debería tener el menor estado energético, es decir, la forma más estable en términos estructurales. La tercera validación (V3) se hizo en la estructura modelo del clúster con mayor representatividad de la dinámica de producción.

Cada método de validación tiene un puntaje con distinto significado y evalúa una distinta propiedad del modelo como son interacciones atómicas, interacciones residuo-residuo, área accesible al solvente, empaquetamiento de los aminoácidos y similitud con estructuras proteicas reales. En la **Figura 2** para el método de validación ERRAT, los valores del factor de calidad para V1 se encuentran en el 85% para los doce complejos. En cambio para V2 y V3 los valores del factor de calidad se encuentran sobre el 90% y en el caso de JAZ1-MYC3, JAZ3-MYC3, JAZ7-MYC3, JAZ8-MYC3 y JAZ12-MYC3 el factor de calidad en la validación V3 sobrepasan el 95%.

En la **Figura 3 y 4** para el método de validación ProQ, V1, V2 y V3 tienen puntajes de LGScore mayor a 1.5 y MaxSub mayor a 0.1. En la **Figura 5** para el método de validación QMean4, V1, V2 y V3 tienen puntajes de alrededor de 0.8 y no existen grandes diferencias entre cada validación. En la **Figura 6 y 7** para el método de validación SolvX, los puntajes de solvatación para V1, V2 y V3 son en su mayoría valores negativos, con ciertas excepciones puntuales como JAZ1 y JAZ2 para la validación V3 y las proteínas MYC3 en el complejo JAZ5-MYC3 y JAZ8-MYC3 en la validación V2 cuyos puntajes de solvatación son positivos.

$\Delta G_{Unión}$ de los complejos JAZ-MYC

Para el cálculo de la energía libre de unión se utilizó las estructuras más prevalentes a lo largo de la dinámica molecular. Para determinar esto se realizó un proceso de clustering basado en el RMSD (Root Mean Square Deviation) de las 5000 estructuras de la trayectoria de producción. En promedio se obtuvo 13 clústeres por complejo JAZ-MYC, cada uno de estos clústeres representan un porcentaje del total de estructuras proteicas de la trayectoria. Se calculó el $\Delta G_{Unión}$ de cada modelo principal de cada clúster. La **Tabla 1**, contiene la descripción de cada clúster y su tamaño para cada complejo JAZ-MYC.

La energía libre de unión en kJ/mol de los modelos finales se encuentra en la **Figura 8**. Todos los valores de energía de unión son negativos lo cual denota acoplamiento energéticamente favorable entre ambas proteínas. El promedio de energía libre de unión de los doce complejos es de -10.94 +/-2.67 kJ/mol. Los promedios de energía libre de cada complejo fueron comparados con dos controles negativos Peapod 1 (PPD1) y Peapod 2 (PPD2) que son proteínas de la misma familia que las proteínas JAZ pero que experimentalmente no tienen la capacidad de unión con MYC3. En la **Figura 8** se puede observar que el valor de la energía de unión de PPD1 es de -5.72 kJ/mol y de PPD2 de -8.32 kJ/mol.

Hotspots complejos JAZ-MYC

Los algoritmos Random Forest, Multilayer Perceptron, Naive Bayes (NB) y Sequential minimal optimization (SMO) fueron analizados para escoger el mejor puntuado en las métricas de calidad. En la **Tabla 3** se encuentran los puntajes obtenidos para las medidas de precisión, exactitud, exhaustividad, puntaje F y Área ROC. Para la selección de atributos se utilizó 3 tipos de algoritmos: RelieveAtribute+Ranker, CorrelationAtributeEval+Ranker y Gain RatioAtribEval+Ranker. El resultado de la reducción de dimensionalidad arrojó 5 de los 20 descriptores energéticos como atributos con mayor poder predictivo. El set de datos de

entrenamiento con los 5 descriptores se encuentra detallado en la **Tabla 2.** Las predicciones de *hotspots* se realizaron únicamente en los residuos de las interfaces proteína-proteína. Las predicciones de *hotspots* para las doce interfaces JAZ-MYC se encuentran en la **Tabla 4.**

Dentro de los *hotspots* predichos para cada interface resaltan residuos aromáticos como fenilalanina e hidrofóbicos como la leucina en el caso de las proteínas JAZ. Dentro de las predicciones para MYC3 se encuentran aminoácidos aromáticos como el triptófano, fenilalanina y aminoácidos con carga negativa como el ácido aspártico y el ácido glutámico. Todas las predicciones se encuentran dentro de la cavidad de interacción JAZ-MYC por lo que se trata de aminoácidos con poca superficie al solvente y empaquetados en la interface.

Discusión

Estructura 3D de los complejos proteicos JAZ-MYC

La generación de las estructuras JAZ-MYC fueron realizadas con modelación por homología. Este método permite predecir estructuras tridimensionales de proteínas con estructura desconocida a partir de plantillas de estructuras 3D conocidas y con cierto nivel de homología estructural. Las proteínas homólogas son consideradas como tal si pertenecen a la misma familia, poseen la misma función y tiene más de un 30% de similitud con la proteína de interés (Vlachakis, 2007). Las proteínas estudiadas cumplen con este requisito pues en el caso de las proteínas JAZ todas pertenecen a la misma familia de proteínas represoras TIFY y comparten varios dominios funcionales. La estructura cristalográfica de MYC3 ya existe por lo que no fue necesario realizar un modelo estructural completo sino únicamente ciertas modificaciones puntuales.

Para la creación de los modelos de las proteínas JAZ se usó como plantilla las proteínas JAZ1 y JAZ9 cuya estructura cristalográfica ya fue obtenida por Zhang et al, 2015. La estructura cristalográfica de MYC3 ya fue obtenida por el mismo autor pero requirió la

adición de aminoácidos faltantes que no se describen en la estructura cristalográfica original pero son importantes dentro de la evaluación de su interacción con las proteínas JAZ. Para esta tarea se utilizaron dos extensiones de Modeller i.e. ModWeb para completar los residuos de MYC3 que no pudieron ser resueltos mediante técnicas de cristalografía y ModLoop para el modelado automatizado de regiones *loop* que no poseen un plegamiento definido (Fiser et al., 2000; Fiser et al., 2003).

Una vez obtenidas las estructuras 3D se requiere evaluar la calidad de los modelos. La validación de la calidad de las estructuras modeladas es importante porque permite determinar qué tan confiable son las predicciones alrededor de un número de parámetros estructurales y energéticos. La calidad de un modelo por homología depende de la distancia evolutiva de la proteína de interés y la plantilla utilizada para su modelado. Como explicamos anteriormente las plantillas utilizadas para el modelado cumplen el criterio de más de 30% de homología con las proteínas a modelar.

Validación de los modelos 3D de los complejos JAZ-MYC

Los servidores de evaluación que se escogieron para la validación fueron Errat, ProQ, QMean4 y SolvX los cuales fueron utilizados en el trabajo de Bhutani et al. 2015 para la validación de modelos estructurales predichos para las proteínas DprE1 y DprE2 que son críticas para la supervivencia de *Mycobacterium tuberculosis.* Cada uno de estos métodos evalúa diferentes características del modelo usando diferentes estrategias y genera puntajes de calidad. Se escogió esta metodología porque DprE1 y DprE2 son proteínas desordenadas y pequeñas, características compartidas con nuestro sistema de estudio.

Se evaluaron los modelos en tres etapas de validación (V1, V2 y V3) con el objetivo de estudiar el progreso de la calidad a lo largo de la dinámica molecular. En general, los modelos obtenidos durante la dinámica final de producción son los mejores puntuados. Sin

embargo, debido a que las predicciones estructurales fueron hechas utilizando como plantillas estructuras cristalográficas con alto porcentaje de homología, los puntajes para V1, V2 y V3 no tuvieron diferencias notables.

Errat evalúa la calidad de las estructuras proteicas utilizando un factor de calidad total que posee una sensibilidad del orden de 1.5Å de error en el esqueleto de la proteína (Colovos & Yeates, 1993). Estructuras con un factor de calidad total de 95% en adelante son considerados buenos modelos (Colovos & Yeates, 1993). Como se observa en la **Figura 2** los valores del factor de calidad para V1 no llegan al estándar de calidad. Sin embargo, a medida que se prolonga el procesamiento por dinámica molecular la calidad mejora. El factor de calidad para V2 y V3 se encuentra sobre el 90% y en el caso de JAZ1-MYC3, JAZ3-MYC3, JAZ7-MYC3, JAZ8-MYC3 y JAZ12-MYC3 el factor de calidad en V3 sobrepasan el 95%.

El método de validación de ProQ tiene dos puntajes de calidad, a saber, LGScore y MaxSub. La diferencia entre ambos indicadores de calidad es su dependencia con el tamaño del sistema. Complejos proteicos de mayor tamaño tienen mayor probabilidad de obtener un mejor LGscore, mientras que complejos pequeños tienden a ser mejor puntuados en MaxSub (Wallner & Elofsson, 2003). En general, modelos estructuralmente correctos poseen LGscore> 1.5 y MaxSub> 0.1 (Wallner & Elofsson, 2003). Como se observa en la **Figura 3 y 4,** la validación V1, V2 y V3 tienen puntajes de LGScore mayor a 1.5 y MaxSub mayor a 0.1.

Es curioso destacar que el puntaje LGScore de la validación V2 es el más alto de todos para los doce complejos. Esto indica que el proceso de minimización de energía resulta favorable para la optimización estructural del complejo. Además el valor de Maxsub para V1 es el más alto en comparación con V2 y V3. Lo que indica que según este método de validación el modelo inicial sin procesamiento de dinámica molecular tiene una estructura mucho más óptima que las estructuras procesadas.

El método de validación QMean4 mide la calidad absoluta de los modelos proteicos y estima el grado de "natividad" estructural del modelo (Benkert et al., 2010). QMean4 es una función de puntuación normalizada independiente del tamaño de la proteína (Benkert et al., 2010). El puntaje de calidad global QMeanDisCo toma en cuenta las métricas de calidad de QMean, las distancias interatómicas y características relacionadas a la estructura primaria de la proteína. QMeanDisCo adquiere valores entre 0 y 1 siendo los puntajes cercanos a 1 los de mayor calidad. Se considera un modelo de alta calidad a los que sobrepasan el puntaje de 0.6 (Benkert et al., 2010). Como se observa en la **Figura 5,** todos los complejos se encuentran en valores superiores a 0.6. La diferencia entre validación V1, V2 y V3 son mínimas lo cual indica que los modelos iniciales fueron de buena calidad.

El método de validación SolvX determina el perfil de solvatación del complejo proteico. Generalmente, valores más negativos son característicos de mejores modelos en proteínas globulares ordenadas. Sin embargo, se espera obtener valores no tan negativos debido a la naturaleza desordenada (no plegada) de los complejos estudiados. Los resultados muestran que los puntajes de solvatación tanto para las proteínas JAZ y MYC3 en su mayoría tienen valores negativos con ciertas excepciones puntuales (**Figura 6 y 7**). Esto puede implicar ciertos problemas estructurales de desnaturalización o desplegamiento de la proteína. Sin embargo, al tratarse de complejos proteicos desordenados, este tipo de comportamiento es habitual por la gran cantidad de regiones *loop* sin estructura fija. Los puntajes de solvatación que no se muestran en la figura no pudieron obtenerse debido a fallas internas en el funcionamiento del servidor relacionadas al formato de las coordenadas de la estructura.

$\Delta G_{unión}$ de los complejos JAZ-MYC

El cálculo de la energía libre de unión entre dos estructuras proteicas es importante para determinar su interacción y rol biológico que tienen dentro de las rutas metabólicas de

las cuales forman parte. Existen varios métodos teóricos para determinar la energía de unión entre complejos proteicos como cálculos MMPBSA que utilizan métodos de dinámica molecular acoplados a cálculos termodinámicos para calcular el ciclo de acoplamiento proteína-proteína (Vreven et al., 2012). Otros métodos computacionalmente menos costosos son el uso de funciones energéticas empíricas que se basan en ensayos experimentales de afinidad (Vreven et al., 2012).

En este trabajo se utilizó la función de energía de FoldX que permite determinar la energía de unión de complejos proteicos en base al cambio de energía libre de Gibbs y a la constante de disociación de cada complejo. Se calculó el ΔG de cada uno de los modelos representantes de cada clúster respectivamente para cada complejo JAZ-MYC. Ensayos experimentales han probado que todas las proteínas JAZ se unen al factor de transcripción MYC3. Sin embargo, por la naturaleza de los ensayos (técnica yeast-two hybrid) es difícil determinar cuáles son las diferencias en términos de afinidad entre cada complejo. Por lo tanto, este enfoque computacional es una excelente alternativa para determinar cuantitativamente esta propiedad energética mediante la predicción de las energías libres de unión.

Como se observa en la **Figura 8**, en promedio la energía de unión de cada complejo se encuentra en -10.94 +/-2.67 kJ/mol. El signo negativo denota que existe liberación de energía y por lo tanto el acoplamiento de los complejos se da espontáneamente lo que concuerda con los ensayos experimentales. Para determinar la validez del método se introdujo dos controles negativos PPD1 y PPD2 que al ser evaluados experimentalmente presentan poca o nula interacción con MYC3.

En la **Figura 8**, se puede observar que el valor de la energía de unión de PPD1 es de -5.72 kJ/mol y de PPD2 de -8.32 kJ/mol. Estos valores difieren del valor promedio de energía de unión de los complejos JAZ (-10.94 +/-2.67 kJ/mol). Estos valores al ser negativos indican

que existe algo de afinidad entre MYC3 y las proteínas PPD. Este resultado es esperado porque las proteínas PPD poseen un dominio Jas muy similar al dominio Jas de las proteínas JAZ. Por otra parte, uno de los clusters de PPD1 presentó una energía libre de unión positiva indicando que esa configuración en particular no se une a MYC3. En general, la única diferencia entre JAZ y PPD se encuentra en un subgrupo de aminoácidos que las proteínas PPD no poseen y que de forma interesante pertenecen al grupo de *Hotspots* predichos por el modelo de machine learning que se desarrolló y que se explicará a continuación.

Hotspots moleculares de los complejos JAZ-MYC3

El set de datos que se recopiló para el entrenamiento del modelo proviene de ensayos experimentales de mutación de sitio dirigido en donde se modificó el aminoácido de interés por alanina y se observó su efecto en la interacción JAZ-MYC3. Se escogieron 20 mutaciones para alimentar el modelo de las cuales 10 son deletéreas (D) y 10 son no deletéreas (ND). Es importante tener un set de datos balanceado entre las dos clases (D y ND) ya que el modelo se vuelve más robusto para el proceso de clasificación en nuevos sets de datos. Además, se excluyó del set de entrenamiento aminoácidos con mutaciones múltiples o mutaciones con otro aminoácido distinto a alanina para tener un set homogéneo.

Los algoritmos Random Forest, Multilayer Perceptron, Naive Bayes (NB) y Sequential minimal optimization (SMO) fueron analizados para escoger el mejor puntuado en las métricas de calidad. Se escogieron estos algoritmos porque han sido evaluados en estudios similares de predicción de *Hotspots* dando buenos resultados (Aguilera-Pesantes et al., 2017; Liu et al., 2009; Wang et al., 2012; Murakami & Mizuguchi, 2010; Mintseris, 2003). Las métricas de evaluación de cada algoritmo determinan la calidad de las predicciones del mismo. En la **Tabla 3** se encuentran los puntajes obtenidos para las medidas de calidad como: precisión, exactitud, exhaustividad, puntaje F y Área ROC. El algoritmo mejor puntuado en dichas métricas fue SMO con un valor de exactitud de 0.84 lo que quiere decir

que el 84% de las predicciones son verdaderos positivos. Por lo tanto, se escogió dicho algoritmo para realizar las predicciones de *hotspots.*

Las predicciones de *hotspots* se realizaron únicamente en los residuos de las interfaces proteína-proteína pues experimentalmente se ha evaluado que esta región es suficiente para generar la interacción. Las predicciones de *hotspots* para las doce interfaces JAZ-MYC se encuentran en la **Tabla 4.** Debido a la homología estructural de estos complejos entre sí, se observa que muchos de los hotspots tanto para las proteínas JAZ como MYC3 son conservados en los doce casos de estudio.

En el caso de las proteínas JAZ, los *hotspots* conservados en las doce estructuras son Serina, Leucina, Fenilalanina y Arginina. Estos aminoácidos se encuentran ubicados en el dominio Jas de las proteínas JAZ y en dirección a los hotspots de su contraparte MYC3. La disposición de estos *hotspots* conservados se mantiene entre complejo y complejo que puede ser indicativo de un motivo corto de reconocimiento molecular de la forma SL••FL•••R. Este motivo sería el principal centro de interacción entre las dos proteínas y el sitio responsable de la mayor afinidad entre moléculas. La principal diferencia entre el motivo Jas de las proteínas PPD1 / PPD2 y el de las proteínas JAZ es la ausencia de los aminoácidos FL y F, respectivamente. De esta manera, se puede confirmar que el motivo corto es esencial para el reconocimiento y acoplamiento de los complejos proteicos.

Los *hotspots* predichos para MYC3 son ricos en aminoácidos aromáticos como el triptófano, fenilalanina y aminoácidos con carga negativa como el ácido aspártico y el ácido glutámico. Principalmente las predicciones caen en la región JID-TAD que constituye el dominio de unión con las proteínas JAZ para su represión y de MED25 para su activación. Las predicciones no presentan un motivo de unión específico que comparten todos los complejos. Sin embargo, los aminoácidos F, L y M son los que más se repiten en las interfaces moleculares JAZ-MYC3. De esta manera, se puede inferir que los *hotspots* de

MYC3 difieren según su acoplamiento con cada una de las 12 proteínas represoras JAZ, lo cual puede ser indicio de distintos mecanismos de represión por parte de cada una de las proteínas JAZ.

Conclusiones

El estudio de la interacción proteína-proteína en las principales rutas de regulación metabólica de plantas y demás organismos vivos es un campo de investigación amplio y con mucho potencial en aplicaciones biotecnológicas. Las herramientas de biología computacional tradicional como alineamiento de secuencia, análisis filogenético y herramientas emergentes como machine learning, dinámica molecular y modelación 3D permiten el estudio a gran escala de datos biológicos.

En el presente estudio se desarrolló un flujo de trabajo computacional en base a datos experimentales que permite obtener tres tipos de predicciones. Primero, la predicción de un modelo estructural 3D válido a partir de secuencias de aminoácidos. Segundo, la predicción de la energía libre de unión para complejos proteína-proteína. Tercero, la predicción de *hotspots* de interacción molecular proteína-proteína.

Todas las predicciones y evaluación de los modelos se realizaron para los doce complejos JAZ-MYC3 que forman parte de la cascada de señalización del ácido jasmónico y son los responsables de la represión o expresión de genes involucrados en la respuesta inmune de la planta. Se logró obtener doce modelos estructurales validados mediante 4 distintos métodos los cuales se utilizaron para las demás etapas del estudio.

Las predicciones de energía de unión promedio para los doce complejos fue de -10.94 +/-2.67 kJ/mol lo cual es indicativo de la existencia de interacción leve característica de interfaces con proteínas desordenadas. Las predicciones de *hotspots* generó resultados interesantes en cuanto a la conservación del motivo corto SL••FL•••R que aparentemente podría tratarse del principal motivo de reconocimiento molecular compartido por todas las interfaces JAZ-MYC.

Recomendaciones

La naturaleza de este trabajo es principalmente computacional, por lo tanto, se requiere de estudios experimentales que validen los resultados obtenidos. Este trabajo de investigación toma como punto de partida estructuras proteicas predichas por homología. Por lo tanto, si en el futuro se logra describir la estructura cristalográfica de todos complejos JAZ-MYC es importante comprobar los resultados obtenidos dentro de este estudio utilizando las estructuras experimentales.

El flujo de trabajo *in silico* descrito en este estudio está bien definido en cuanto a la metodología utilizada. Sin embargo, para poder reproducirlo de forma más fácil con cualquier otro sistema proteico es necesario realizar una implementación de todos los pasos del flujo de trabajo en una sola aplicación. Esto permite reducir errores y el tiempo empleado en reproducir la metodología. Para este efecto, se requieren conocimientos en ingeniería de sistemas que están fuera del área de conocimiento de la carrera de biotecnología.

Referencias bibliográficas

Aguilera-Pesantes, D., Robayo, L. E., Méndez, P. E., Mollocana, D., Marrero-Ponce, Y., Torres, F. J., & Méndez, M. A. (2017). Discovering key residues of dengue virus NS2b-NS3-protease: New binding sites for antiviral inhibitors design. *Biochemical and Biophysical Research Communications*, 492(4), 631-642.

Bhutani, I., Loharch, S., Gupta, P., Madathil, R., & Parkesh, R. (2015). Structure, dynamics, and interaction of Mycobacterium tuberculosis (Mtb) DprE1 and DprE2 examined by molecular modeling, simulation, and electrostatic studies. *PloS one*, 10(3), e0119771.

Bienert, S., Waterhouse, A., de Beer, T. A., Tauriello, G., Studer, G., Bordoli, L., & Schwede, T. (2016). The SWISS-MODEL Repository—new features and functionality. *Nucleic acids research*, 45(D1), D313-D319.

Bogan, A. A., & Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. *Journal of molecular biology*, 280(1), 1-9.

Causier, B., & Davies, B. (2002). Analysing protein-protein interactions with the yeast two-hybrid system. *Plant molecular biology*, 50(6), 855-870.

Colovos, C., & Yeates, T. O. (1993). Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Science : A Publication of the Protein Society*, *2*(9), 1511–1519.

Fiser, A., & Do, R. K. G. (2000). Modeling of loops in protein structures. *Protein science*, *9*(9), 1753-1773.

Fiser, A., & Sali, A. (2003). ModLoop: automated modeling of loops in protein structures. *Bioinformatics*, *19*(18), 2500-2501.

Frank, E., Hall, M., Trigg, L., Holmes, G., & Witten, I. H. (2004). Data mining in bioinformatics using Weka. *Bioinformatics*, 20(15), 2479-2481.

Holm, L., & Sander, C. (1992). Evaluation of protein models by atomic solvation preference. *Journal of molecular biology*, *225*(1), 93-105.

Katsir, L., Chung, H. S., Koo, A. J., & Howe, G. A. (2008). Jasmonate signaling: a conserved mechanism of hormone sensing. *Current opinion in plant biology*, 11(4), 428-435.

Kazan, K., & Manners, J. M. (2012). JAZ repressors and the orchestration of phytohormone crosstalk. *Trends in plant science*, 17(1), 22-31.

Kenneth Morrow, J., & Zhang, S. (2012). Computational prediction of protein hot spot residues. *Current pharmaceutical design*, *18*(9), 1255-1265.

Liu, B., Wang, X., Lin, L., Tang, B., Dong, Q., & Wang, X. (2009). Prediction of protein binding sites in protein structures using hidden Markov support vector machine. *BMC bioinformatics*, 10(1), 381.

Lovell, S. C., Davis, I. W., Arendall, W. B., de Bakker, P. I., Word, J. M., Prisant, M. G., ... & Richardson, D. C. (2003). Structure validation by Cα geometry: ϕ, ψ and Cβ deviation. *Proteins: Structure, Function, and Bioinformatics*, *50*(3), 437-450.

Memelink, J. (2009). Regulation of gene expression by jasmonate hormones. *Phytochemistry*, 70(13), 1560-1570.

Mintseris, J., & Weng, Z. (2003). Atomic contact vectors in protein-protein recognition. *Proteins: Structure, Function, and Bioinformatics*, 53(3), 629-639.

Moreira, I. S., Fernandes, P. A., & Ramos, M. J. (2007). Hot spots—A review of the protein–protein interface determinant amino-acid residues. *Proteins: Structure, Function, and Bioinformatics*, 68(4), 803-812.

Murakami, Y., & Mizuguchi, K. (2010). Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites. Bioinformatics, 26(15), 1841-1848.

Pauwels, L., & Goossens, A. (2011). The JAZ proteins: a crucial interface in the jasmonate signaling cascade. *The Plant Cell*, 23(9), 3089-3100.

Ramachandran, G. N., Ramakrishnan, C., & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Journal of molecular biology*, *7*(1), 95-99.

Santner, A., & Estelle, M. (2007). The JAZ proteins link jasmonate perception with transcriptional changes. *The Plant Cell*, 19(12), 3839-3842.

Schweizer, F., Fernández-Calvo, P., Zander, M., Diez-Diaz, M., Fonseca, S., Glauser, G., ... & Reymond, P. (2013). Arabidopsis basic helix-loop-helix transcription factors MYC2, MYC3, and MYC4 regulate glucosinolate biosynthesis, insect performance, and feeding behavior. *The Plant Cell*, 25(8), 3117-3132.

Schmidt, T., Bergner, A., & Schwede, T. (2014). Modelling three-dimensional protein structures for applications in drug design. *Drug discovery today*, 19(7), 890-897.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic acids research*, *33*(suppl_2), W382-W388.

Staswick, P. E. (2008). JAZing up jasmonate signaling. *Trends in plant science*, 13(2), 66-71

The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC

Vasyukova, N. I., & Ozeretskovskaya, O. L. (2009). Jasmonate-dependent defense signaling in plant tissues. *Russian journal of plant physiology*, 56(5), 581-590.

Vlachakis, D. (2007). *An Introduction to Molecular Modelling, from theory to application*. Lulu. com.

Vreven, T., Hwang, H., Pierce, B. G., & Weng, Z. (2012). Prediction of protein–protein binding free energies. *Protein Science*, 21(3), 396-404.

Wallner, B., & Elofsson, A. (2003). Can correct protein models be identified? *Protein science*, *12*(5), 1073-1086.

Wang, L., Liu, Z. P., Zhang, X. S., & Chen, L. (2012). Prediction of hot spots in protein interfaces using a random forest model with hybrid features. *Protein Engineering, Design & Selection*, 25(3), 119-126.

Wasternack, C., & Hause, B. (2013). Jasmonates: biosynthesis, perception, signal transduction and action in plant stress response, growth and development. An update to the 2007 review in Annals of Botany. *Annals of botany*, 111(6), 1021-1058.

Webb, B., & Sali, A. (2014). Protein structure modeling with MODELLER. *Protein Structure Prediction*, 1-15.

Zhang, F., Yao, J., Ke, J., Zhang, L., Lam, V. Q., Xin, X. F., ... & Zhou, M. (2015). Structural basis of JAZ repression of MYC transcription factors in jasmonate signaling. *Nature*, 525(7568), 269.

ANEXOS

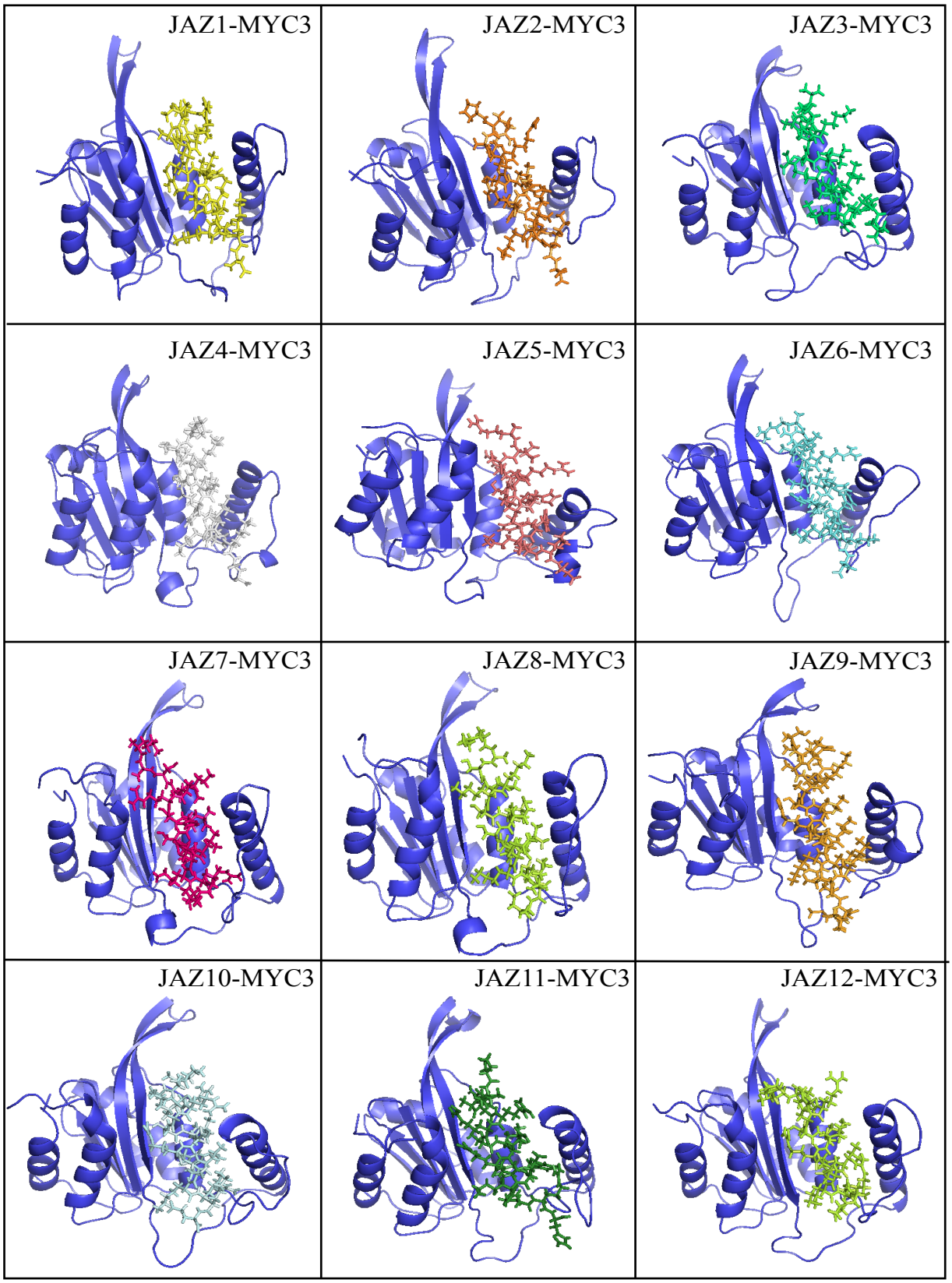
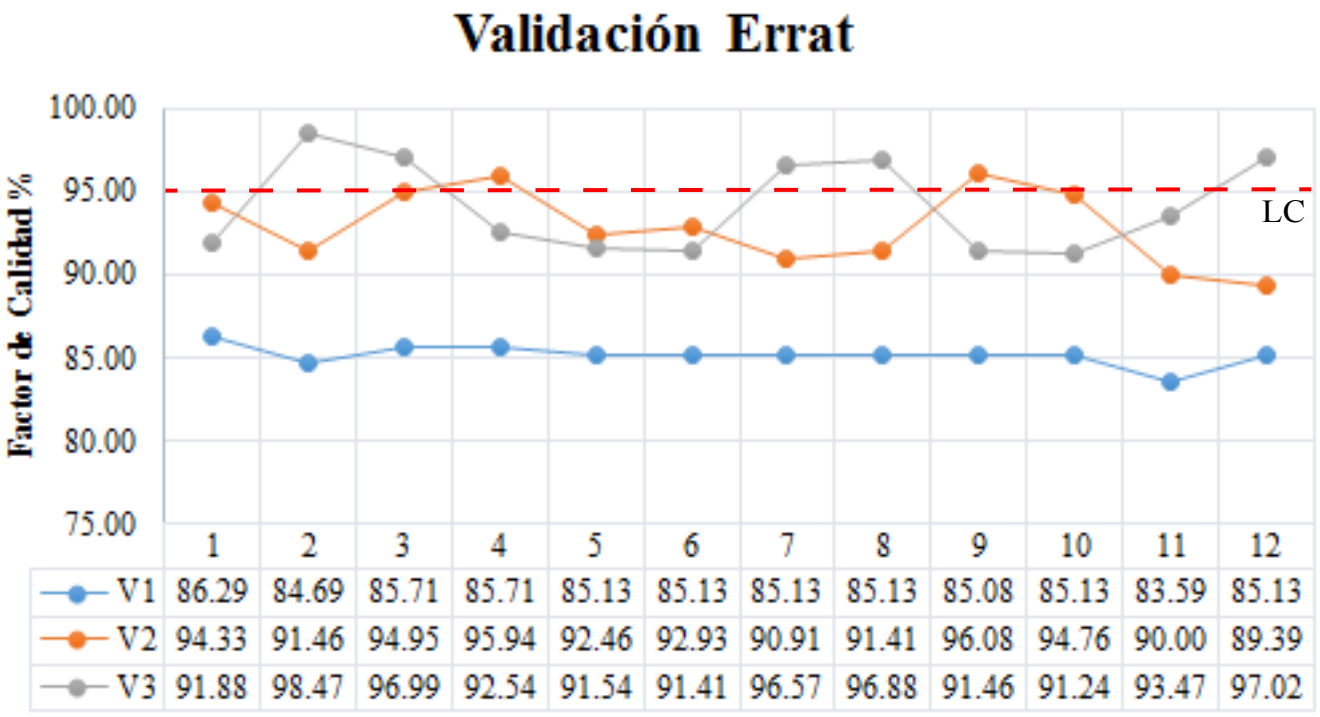
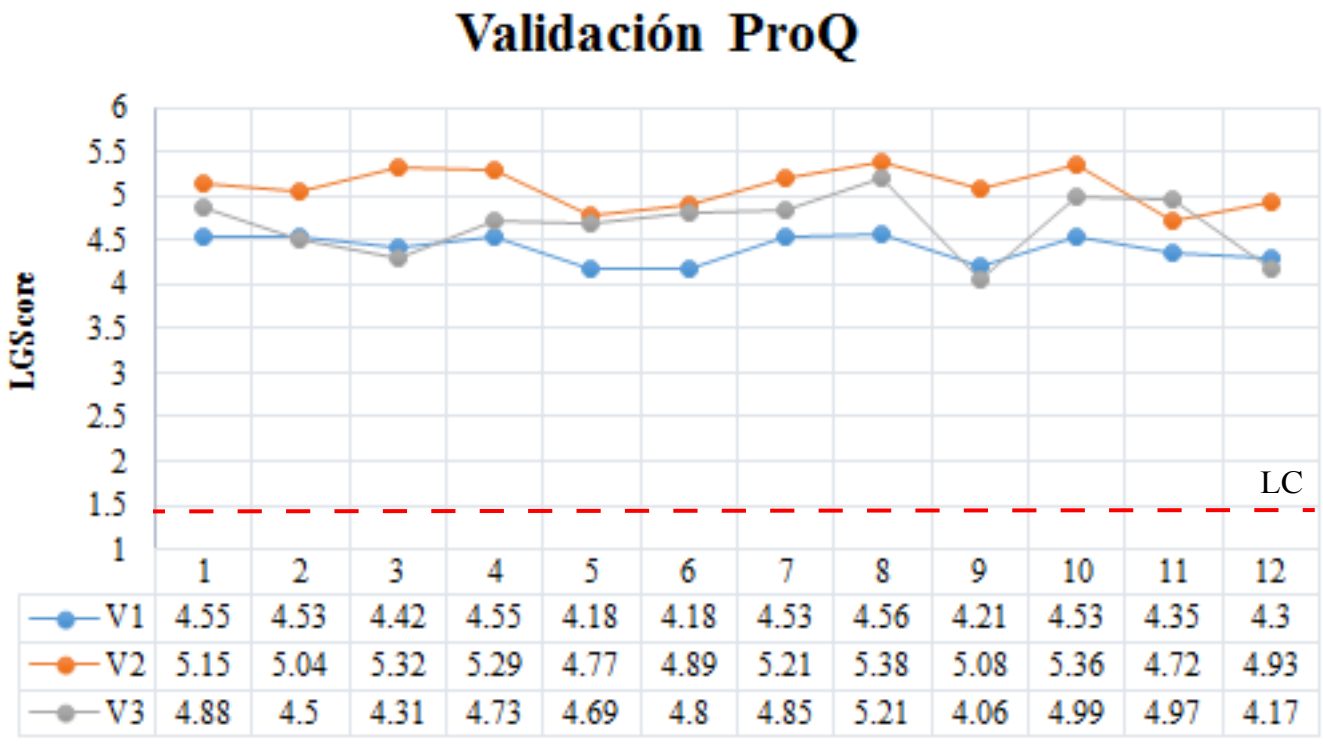


Figura 1. Modelos estructurales validados de los doce complejos proteicos JAZ-MYC3. La representación gráfica de los modelos se hizo en Pymol 1.7.4.5. La proteína MYC3 se encuentra en representación *cartoon* y color azul para todos los complejos. Las proteínas JAZ se encuentran en representación *sticks* y en diferentes colores en cada uno de los complejos.



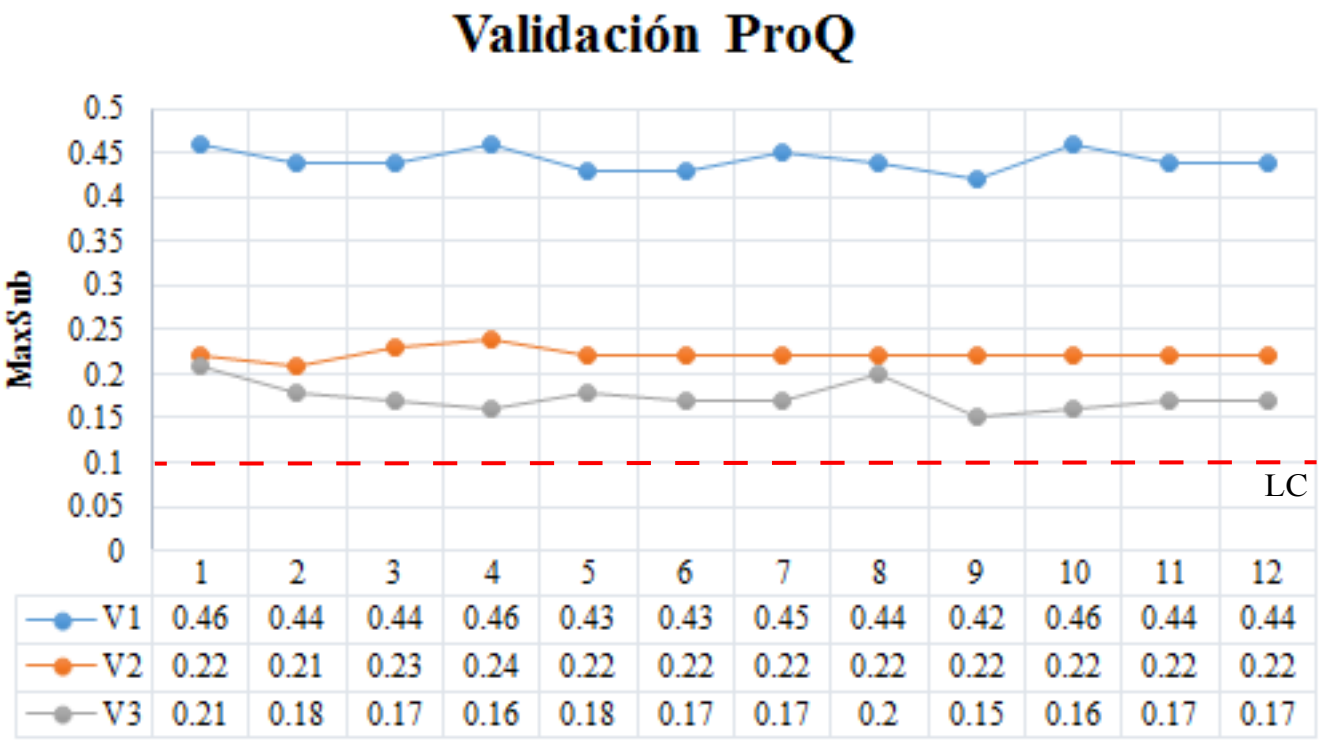
	1	2	3	4	5	6	7	8	9	10	11	12
V1	86.29	84.69	85.71	85.71	85.13	85.13	85.13	85.13	85.08	85.13	83.59	85.13
V2	94.33	91.46	94.95	95.94	92.46	92.93	90.91	91.41	96.08	94.76	90.00	89.39
V3	91.88	98.47	96.99	92.54	91.54	91.41	96.57	96.88	91.46	91.24	93.47	97.02

Figura 2. Validación cuantitativa de los modelos estructurales de los doce complejos proteicos JAZ-MYC3 utilizando el protocolo de validación de Errat. El gráfico describe el factor de calidad (%) en el eje *y* y el número de complejo en el eje *x*. La línea azul representa la primera validación (V1), la línea naranja representa la segunda validación (V2) y la línea gris representa la tercera validación (V3). La línea roja entrecortada indica el límite de calidad (LC), valores sobre el LC indican modelos de alta calidad.



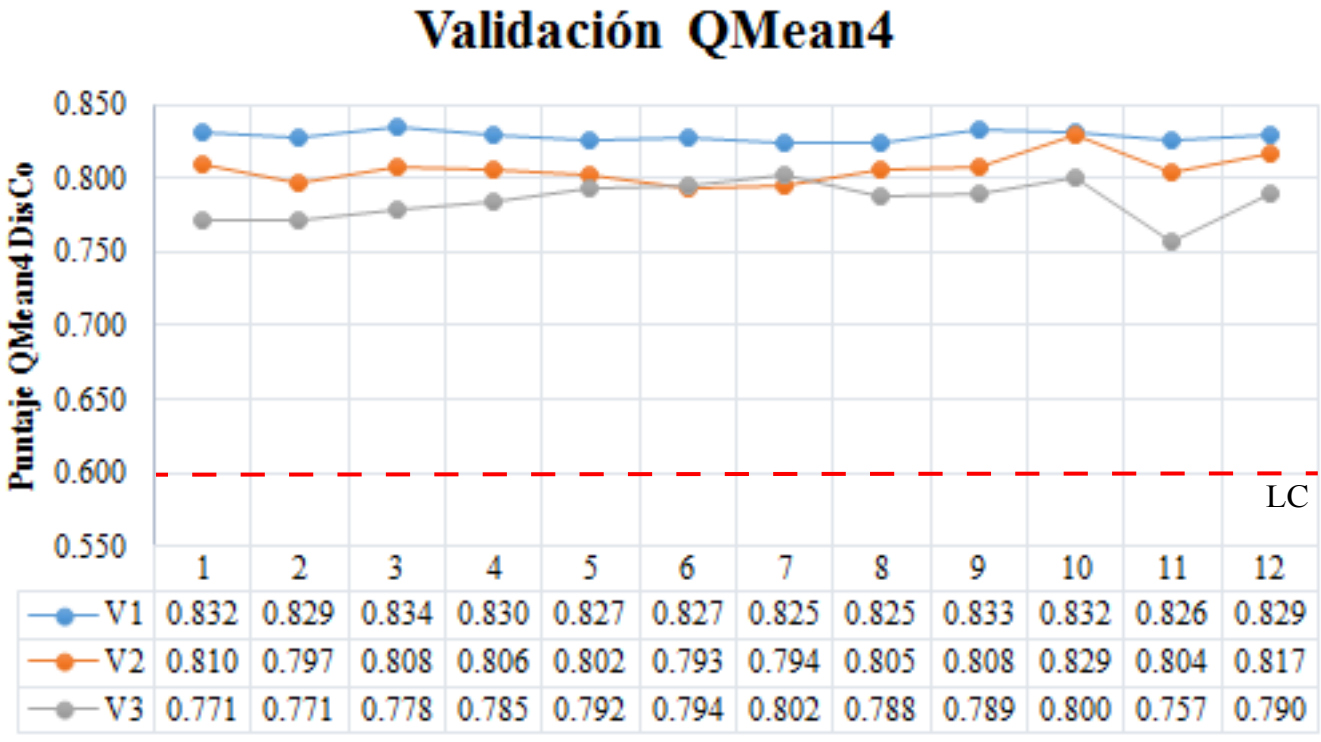
	1	2	3	4	5	6	7	8	9	10	11	12
V1	4.55	4.53	4.42	4.55	4.18	4.18	4.53	4.56	4.21	4.53	4.35	4.3
V2	5.15	5.04	5.32	5.29	4.77	4.89	5.21	5.38	5.08	5.36	4.72	4.93
V3	4.88	4.5	4.31	4.73	4.69	4.8	4.85	5.21	4.06	4.99	4.97	4.17

Figura 3. Validación cuantitativa de los modelos estructurales de los doce complejos proteicos JAZ-MYC3 utilizando el protocolo de validación de ProQ. El gráfico describe el LGScore en el eje *y* y el número de complejo en el eje *x*. La línea azul representa la primera validación (V1), la línea naranja representa la segunda validación (V2) y la línea gris representa la tercera validación (V3). La línea roja entrecortada indica el límite de calidad (LC), valores sobre el LC indican modelos de alta calidad.



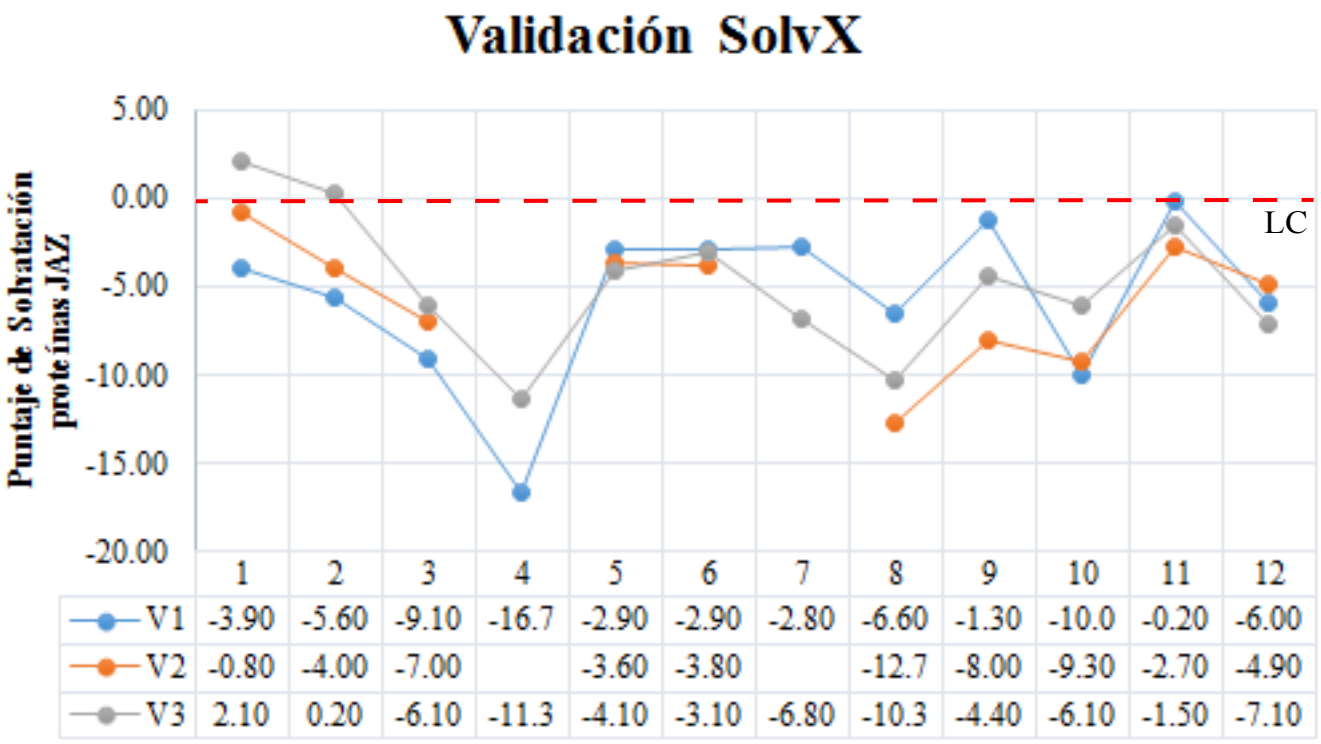
	1	2	3	4	5	6	7	8	9	10	11	12
V1	0.46	0.44	0.44	0.46	0.43	0.43	0.45	0.44	0.42	0.46	0.44	0.44
V2	0.22	0.21	0.23	0.24	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22
V3	0.21	0.18	0.17	0.16	0.18	0.17	0.17	0.2	0.15	0.16	0.17	0.17

Figura 4. Validación cuantitativa de los modelos estructurales de los doce complejos proteicos JAZ-MYC3 utilizando el protocolo de validación de ProQ. El gráfico describe el MaxSub en el eje *y* y el número de complejo en el eje *x*. La línea azul representa la primera validación (V1), la línea naranja representa la segunda validación (V2) y la línea gris representa la tercera validación (V3). La línea roja entrecortada indica el límite de calidad (LC), valores sobre el LC indican modelos de alta calidad.



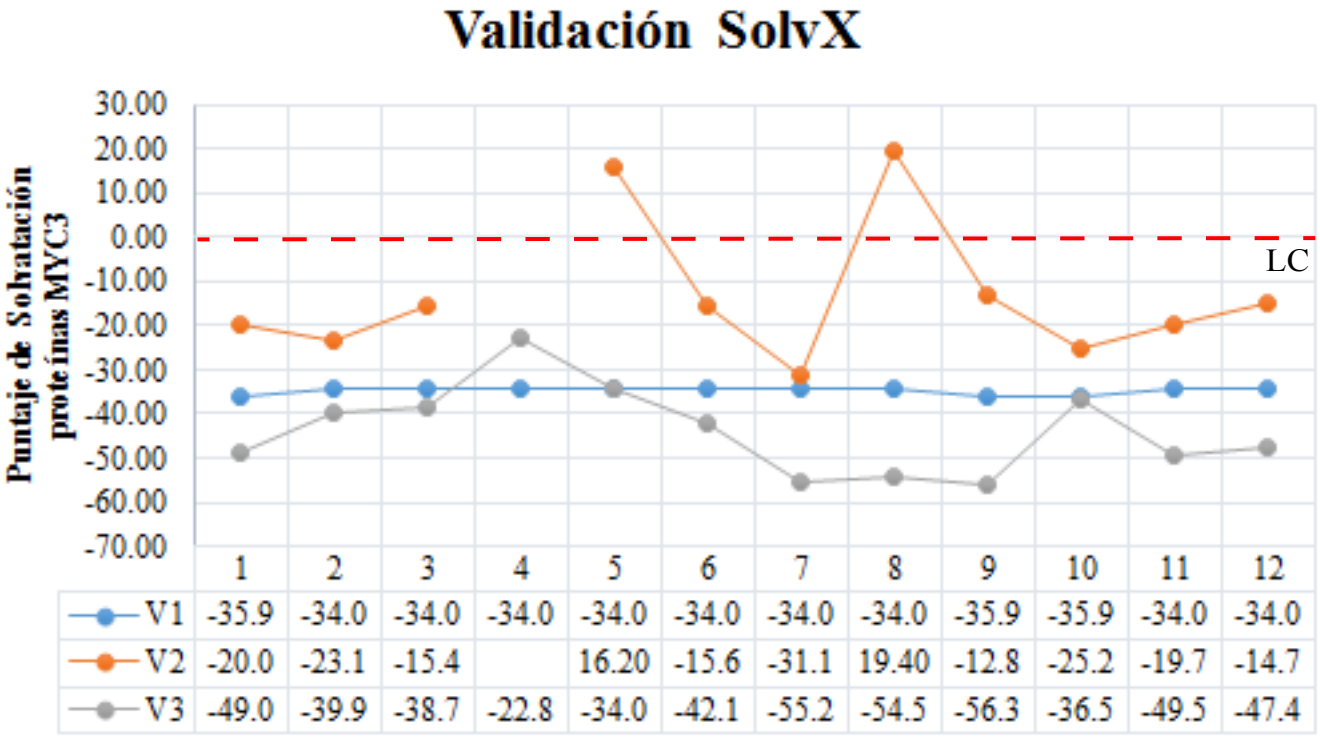
	1	2	3	4	5	6	7	8	9	10	11	12
V1	0.832	0.829	0.834	0.830	0.827	0.827	0.825	0.825	0.833	0.832	0.826	0.829
V2	0.810	0.797	0.808	0.806	0.802	0.793	0.794	0.805	0.808	0.829	0.804	0.817
V3	0.771	0.771	0.778	0.785	0.792	0.794	0.802	0.788	0.789	0.800	0.757	0.790

Figura 5. Validación cuantitativa de los modelos estructurales de los doce complejos proteicos JAZ-MYC3 utilizando el protocolo de validación de Qmean4. El gráfico describe el LGScore en el eje *y* y el número de complejo en el eje *x*. La línea azul representa la primera validación (V1), la línea naranja representa la segunda validación (V2) y la línea gris representa la tercera validación (V3), valores sobre el LC indican modelos de alta calidad.



	1	2	3	4	5	6	7	8	9	10	11	12
V1	-3.90	-5.60	-9.10	-16.7	-2.90	-2.90	-2.80	-6.60	-1.30	-10.0	-0.20	-6.00
V2	-0.80	-4.00	-7.00		-3.60	-3.80		-12.7	-8.00	-9.30	-2.70	-4.90
V3	2.10	0.20	-6.10	-11.3	-4.10	-3.10	-6.80	-10.3	-4.40	-6.10	-1.50	-7.10

Figura 6. Validación cuantitativa de los modelos estructurales de los doce complejos proteicos JAZ-MYC3 utilizando el protocolo de validación de SolvX. El gráfico describe el puntaje de solvatación para las proteínas JAZ en el eje *y* y el número de complejo en el eje *x*. La línea azul representa la primera validación (V1), la línea naranja representa la segunda validación (V2) y la línea gris representa la tercera validación (V3). La línea roja entrecortada indica el límite de calidad (LC), valores bajo el LC indican modelos de alta calidad.



	1	2	3	4	5	6	7	8	9	10	11	12
V1	-35.9	-34.0	-34.0	-34.0	-34.0	-34.0	-34.0	-34.0	-35.9	-35.9	-34.0	-34.0
V2	-20.0	-23.1	-15.4		16.20	-15.6	-31.1	19.40	-12.8	-25.2	-19.7	-14.7
V3	-49.0	-39.9	-38.7	-22.8	-34.0	-42.1	-55.2	-54.5	-56.3	-36.5	-49.5	-47.4

Figura 7. Validación cuantitativa de los modelos estructurales de los doce complejos proteicos JAZ-MYC3 utilizando el protocolo de validación de SolvX. El gráfico describe el puntaje de solvatación para las proteínas MYC3 en el eje *y* y el número de complejo en el eje *x*. La línea azul representa la primera validación (V1), la línea naranja representa la segunda validación (V2) y la línea gris representa la tercera validación (V3). La línea roja entrecortada indica el límite de calidad (LC),valores bajo el LC indican modelos de alta calidad.

Tabla1. Detalles del proceso de clusterizado de los complejos JAZ-MYC3.

COMPLEJO	# CLÚSTERES	TAMAÑO CLÚSTER 1	PORCENTAJE CLÚSTER 1
JAZ1MYC3	9	3238	64.75%
JAZ2MYC3	11	3907	78.12%
JAZ3MYC3	10	3876	77.50%
JAZ4MYC3	13	3137	62.72%
JAZ5MYC3	11	2998	59.94%
JAZ6MYC3	14	2399	47.97%
JAZ7MYC3	16	2622	52.43%
JAZ8MYC3	12	3524	70.47%
JAZ9MYC3	14	2962	59.22%
JAZ10MYC3	18	2640	52.79%
JAZ11MYC3	20	1505	30.09%
JAZ12MYC3	11	3444	68.86%

Interpretación: La tabla muestra el número de clústers que se obtuvieron de cada uno de las simulaciones por dinámica molecular de cada complejo JAZ-MYC3. Se detalla el tamaño del clúster más representativo y su porcetaje de representatividad a lo largo de la dinámica de producción de 50ns (5000 estructuras).

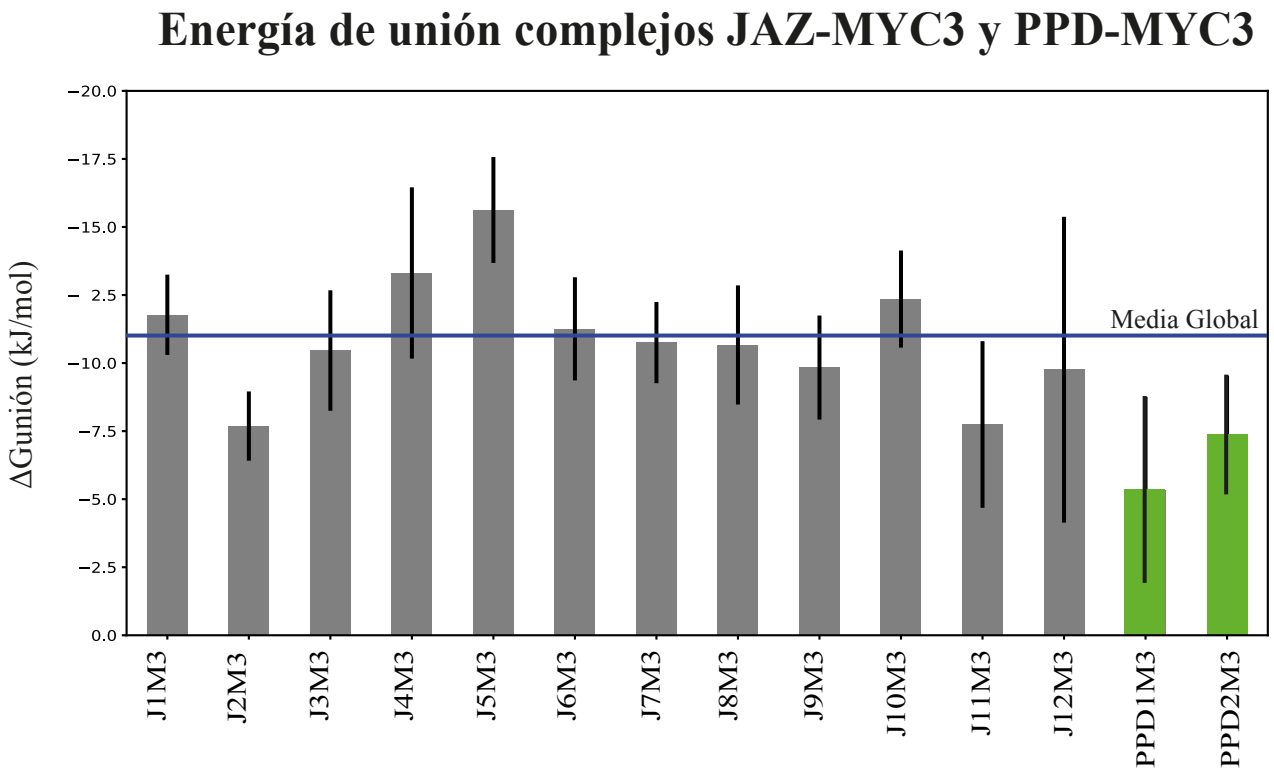


Figura 8. Energía libre de unión en kJ/mol de los complejos JAZ-MYC3 y PPD-MYC3. El gráfico describe la energía libre de unión en el eje *y* y el nombre del complejo en el eje *x*. Las barras grises representan los doce complejos JAZ. Las barrras verdes los controles negativos de los dos complejos PPD1-MYC3 y PPD2-MYC3. La línea azul indica la media global de los complejos JAZ-MYC3 (-10.94 +/-2.67 kJ/mol). Las lineas negras de cada barra son el error estándar.

Tabla 2. Set de datos de entrenamiento con mutaciones de sitio dirigido para el modelo de aprendizaje de máquina.

SET DE RESIDUOS EXPERIMENTALES				MUTAGENESIS DE SITIO DIRIGIDO *IN SILICO* (ΔΔG kJ/mol)					
Clase	**Residuo**	**Tipo de Residuo**	**Referencia**	**A**	**I**	**K**	**P**	**S**	**V**
ND	M3:92	TRP	Zhang et al., 2015	1.3	0.55	0.36	0.53	1.32	0.31
D	M3:94	ASP	Smolen et al., 2002; Zhang et al., 2015	1.53	1.22	1.88	2.1	1.52	1.78
ND	M3:97	TYR	Zhang et al., 2017; Zhang et al., 2015	0.59	-0.38	0.4	-0.52	0.72	-0.42
ND	M3:120	ARG	Zhang et al., 2017	-0.05	-0.04	-0.03	-0.05	-0.04	0
ND	M3:148	GLU	Zhang et al., 2017	0.65	0.36	1.51	1.09	0.75	0.66
D	M3:151	PHE	Zhang et al., 2015	0.73	0.56	1.04	0.7	0.7	0.58
D	M3:152	LEU	Zhang et al., 2015	1.11	1.22	1.24	2.24	1.5	1.17
D	M3:155	MET	Zhang et al., 2015; Zhang et al., 2017	2.11	1.3	1.97	2.42	2.33	2.02
ND	M3:156	THR	Zhang et al., 2017	-0.22	-0.27	-0.14	-0.31	-0.05	-0.16
ND	J9:223	ARG	Melotto et al., 2008	0.52	-0.43	0.38	1.03	-0.05	-0.4
ND	J9:224	LYS	Melotto et al., 2008	1.25	-0.66	-0.24	1.36	1.4	0.16
ND	J9:226	SER	Zhang et al., 2015	0.007	0.33	0.76	1.64	0.01	0.53
D	J9:227	LEU	Zhang et al., 2015	2.94	2	1.61	4.3	4.09	2.24
D	J9:230	PHE	Zhang et al., 2015	3.57	2.29	2.83	5.32	4.26	2.05
D	J9:231	LEU	Zhang et al., 2015	3.31	1.41	3.38	3.16	2.87	1.23
ND	J9:233	LYS	Withers et al., 2012	0.22	0.3	-0.07	0.13	0.13	0.09
D	J9:234	ARG	Withers et al., 2012; Zhang et al., 2015	1.73	0.78	0.53	1.02	1.44	1.38
ND	J9:235	LYS	Withers et al., 2012	0.5	0.05	-0.12	1.14	0.49	0.88
D	J1:205	ARG	Melotto et al., 2008	0.61	0.03	0.6	0.94	1.05	-0.49
D	J1:206	ARG	Melotto et al., 2008	0.46	0.53	-0.02	0.76	0.5	0.13

Interpretación: La tabla muestra la descripción de cada instancia del set de datos de entrenamiento. La clase D, indica que el residuo al ser mutado a Alanina es deletéreo para la interacción JAZ-MYC3. La clase ND, indica que el residuo al ser mutado a Alanina no es deletéreo para la interacción JAZ-MYC3. El set de datos fue obtenido de estudios experimentales de mutación de sitio dirigido de los complejos JAZ1-MYC3 y JAZ9-MYC3. Los valores a la izquierda de la tabla indican el ΔΔG en kJ/mol al mutar el residuo a uno de los aminoácidos A, I, K, P, S, V.

Tabla 3. Evaluación de la ejecución de los algoritmos aprendizaje de máquina Random Forest (RF), Multilayer Perceptron (MLP), Naive Bayes (NB) y Sequential minimal optimization (SMO).

ALGORITMO	PRECISIÓN	EXHAUSTIVIDAD	EXACTITUD	PUNTAJE F	ÁREA ROC
RF	0.74	0.74	0.74	0.74	0.78
SMO	0.88	0.84	0.84	0.84	0.85
MLP	0.58	0.58	0.58	0.58	0.64
NB	0.81	0.79	0.79	0.79	0.81

Interpretación: La tabla muestra las métricas de calidad de cada uno de los algoritmos evaluados para el desarrollo del modelo de aprendizaje automático. Los puntajes van del 0 a 1, mientras más cercano a 1 es el puntaje de cada métrica mejor es el algoritmo.

Tabla 4. Predicciones de *hotspots* de los doce complejos JAZ-MYC3 en clústeres.

COMPLEJO	HOTSPOTS JAZ	HOTSPOTS MYC3
JAZ1MYC3	R:205, R:206, S:208, L:209, F:212, L:213, R:216	I:122, I:129, D:94, F:151, L:152, M:155
JAZ2MYC3	S:210, L:211, F:214, L:215, R:218	W:92, D:94, I:122, I:129, E:148, F:151, M:155
JAZ3MYC3	R:205, L:209, F:212, L:213, R:216	W:92, L:125, I:129, S:136, E:148, M:155
JAZ4MYC3	S:265, L:266, R:268, F:269, L:270, R:273	W:92, D:94, I:122, I:129, E:148, F:151, L:152, M:155
JAZ5MYC3	R:284, S:287, L:288, F:291, F:292, R:295	W:92, D:94, L:125, N:126, I:129, D:137, N:140, E:148, L:152, M:155
JAZ6MYC3	R:285, S:291, L:292, F:295, F:296, K:298, R:299, K:300	W:92, D:94, I:122, N:126, I:129, D:137, E:148, F:151, L:152, M:155
JAZ7MYC3	R:225, S:226, L:227, F:230, L:231, K:233, R:234	I:122, I:129, D:137, S:139, F:151, M:155
JAZ8MYC3	K:208, S:209, L:210, F:213, K:216, R:217	W:92, I:129, N:140, D:141, V:144, E:148, F:151, L:152, M:155
JAZ9MYC3	R:237, L:227, F:230, L:231, R:234	L:125, I:129, D:94, F:151, L:152, M:155
JAZ10MYC3	S:226, L:227, F:230, L:231, R:234, K:235	W:92, D:94, I:122, L:125, E:148, F:151, L:152, M:155
JAZ11MYC3	S:303, L:304, R:306, F:307, F:308, R:311, R:312	D:50, D:94, L:125, I:129, S:130, E:138, E:148, F:151, L:152, M:155
JAZ12MYC3	R:243, S:245, L:246, R:248, F:249, L:250, R:253, R:254	W:92, D:94, I:129, D:141, E:148, L:152, M:155

Interpretación: La tabla muestra los resultados de las predicciones de *hotspots* para los complejos JAZ-MYC3. Cada residuo está nombrado según la naturaleza del aminoácido y su posición en la secuencia. Los residuos marcados en rojo son los que estuvieron dentro del set de entrenamiento.