Universidad Nacional de San Martín

Identificación sistemática de motivos lineales de interacción con la familia de proteínas *pocket*

Proyecto final para optar por el título de Licenciada en Biotecnología
de la Universidad Nacional de San Martín

Carla Lorenze

Directora: Dra. Lucía B. Chemes

Co-Directora: Dra. Juliana Glavina

Lugar de trabajo: Instituto de Investigaciones Biotecnológicas

San Martín, 25 de noviembre de 2024

Índice

Agradecimientos

Agradezco a la Universidad de San Martín (UNSAM) , mi segunda casa durante años y donde me formé; al Instituto de Investigaciones Biotecnológicas (IIB) por darme un espacio de exploración e incentivar mi curiosidad; a cada integrante del Laboratorio de Estructura, Plasticidad y Función de Proteínas por abrirme las puertas y guiarme en el último tramo de mi aprendizaje.

También agradezco profundamente a mis padres y hermano, Norma, Jorge y Lucas, que sin su apoyo incondicional, no hubiera sido lo mismo. A mi pareja, Miguel, quien supo contener y acompañar los altibajos. A todos los amigos y colegas que me llevo de tantos años y experiencias compartidas. Y por último y no menos importante, a mi mascota Menta Granizada, que me acompañó día y noche en todos estos años de estudio.

Glosario

Alineamiento múltiple de secuencia.	Alineamiento de tres o más secuencias de proteínas, ADN o ARN
Bolsillo de unión a LxCxE.	Sitio de unión a través del cual interactúan proteínas que contienen al SLiM LxCxE presente en el dominio pocket
Core.	Residuos del SLiM que establecen contacto directo con la superficie de un dominio globular.
Docking.	Función de acoplamiento de complejos que desempeñan algunos tipos de SLiMs
Dominio globular.	Segmento modular proteico que adquiere estructura terciaria estable
Dominio intrínsecamente desordenado.	Segmento modular proteico que no adquiere estructura terciaria estable
Dominio pocket.	Dominio globular central altamente conservado en las proteínas de la familia pocket compuesto por el subdominio A y el subdominio B
E2F-like.	Variante novedosa del SLiM E2F
Expresión regular.	Patrón de caracteres que representa una secuencia que puede ser identificada en una cadena de texto
Familia Pocket.	Familia de proteínas parálogas que presentan un dominio central 'pocket' altamente conservado que incluyen a retinoblastoma, p107 y p130
Hendidura AB.	Interfaz entre los subdominios A y B del dominio pocket que se encuentra altamente conservado y es el sitio de unión de interactores que contienen el SLiM E2F
Linker.	Segmentos aminoacídicos que generalmente conectan dominios globulares
Logo de secuencia.	Gráfico construido a partir de un alineamiento múltiple de secuencias que consiste en letras apiladas para cada posición donde el tamaño de cada letra es directamente proporcional a la frecuencia de cada letra en cada posición y la altura total corresponde a la conservación de la posición
Loop.	Segmento aminoacídico flexible que puede contener SLiMs de interacción

Loop AB.	Segmento aminoacídico flexible que conecta los subdominios A y B del dominio pocket
Missing residues.	Residuos incluidos en un experimento de cristalografía de rayos X y que no pudieron ser asignados en la estructura obtenida.
Módulo de interacción.	Segmentos proteicos que median interacciones proteína-proteína
Negative Binders.	Interactores hit del ensayo ProP-PD que fueron ensayados experimentalmente en los que no se detectó interacción con los dominios pocket
Péptido hit.	Péptidos de 16 residuos fusionados a la cápside de fagos que fueron aciertos o *hits* en las rondas de selección del ensayo ProP-PD
Phage Display.	Técnica de display de péptidos en la superficie de fagos utilizada para estudiar interacciones proteína-proteína
Posición fija.	Residuos que determinan la unión con un dominio proteico estableciendo contacto directo con la superficie
Posición flanqueante.	Residuos que pueden modular la afinidad de un SLiM por el dominio globular estableciendo interacciones variables y dinámicas con el dominio. Se ubican en los extremos N- y C- terminal del core del SLiM
Posición variable.	Residuos que no interactúan con la superficie del dominio globular y se orientan hacia afuera de la unión
Proteína intrínsecamente desordenada.	Proteínas que presentan cierto grado de desorden en su secuencia. Contienen segmentos ordenados o estructurados y segmentos desordenados o sin estructura
Recall.	Métrica de calidad establecida para estimar la calidad del ensayo ProP-PD
Región intrínsecamente desordenada.	Segmentos proteicos sin estructura terciaria en los que se encuentran los dominios intrínsecamente desordenados y motivos lineales o SLiMs
Short Linear Motif.	Elementos modulares pequeños de secuencia que se ubican en regiones intrínsecamente desordenadas. También conocidos como motivos lineales
Strong Positive Binders.	Interactores hit del ensayo ProP-PD que fueron ensayados experimentalmente en los que se detectó una interacción fuerte con los dominios pocket

Técnicas de gran escala.	Técnica de laboratorio en la que pueden procesarse un alto número de muestras en un mismo ensayo
Técnicas de pequeña escala.	Técnica de laboratorio que utiliza pequeñas cantidades de muestra para obtener resultados
Weak Binders.	Interactores hit del ensayo ProP-PD que fueron ensayados experimentalmente en los que se detectó una interacción débil con los dominios pocket

Abreviaturas

CASP: Critical Assessment ot techniques for protein Structure Prediction

CDK: Quinasa dependiente de Ciclina

$\Delta\Delta G$**:** Variación de Energía Libre de Gibbs

ELM: Eukaryotic Linear Motif Database

HD2: Human Desorderome 2

IDD: Dominio Intrínsecamente Desordenado

IDP: Proteína Intrínsecamente Desordenada

IDR: Región Intrínsecamente Desordenada

MSA: Alineamiento Múltiple de Secuencia

NES: Nuclear Export Signal

NGS: Next Generation Sequencing

NLS: Nuclear Localization Signal

POSIX: Portable Operating System Interface for Unix

ProP-PD: Proteomic Peptide Phage Display

PTM: Modificación Post Traduccional

Rb: Retinoblastoma

RBC: Extremo C-terminal de la proteína Rb

RBN: Extremo N-terminal de la proteína Rb

Regex: Expresión Regular (Regular Expression)

RSA: Accesibilidad Relativa al Solvente

SLiM: Short Linear Motif

TN: True Negative

TP: True Positive

Capítulo 1: Introducción

1.1. Interacciones proteína-proteína

La célula es un sistema dinámico en el cual los procesos fundamentales están mediados por interacciones moleculares, que llevan a cabo funciones que incluyen desde la señalización intracelular hasta la regulación de la expresión génica. Muchas de estas funciones son realizadas por proteínas.

Las interacciones entre proteínas están mediadas por módulos de interacción, siendo los dominios globulares los más estudiados. Estos responden al paradigma clásico de secuencia-estructura-función que sostiene que la secuencia de aminoácidos de la proteína define una estructura tridimensional estable en el tiempo la cual le confiere una función biológica [1].

En los últimos años este paradigma fue modificado por la identificación de regiones proteicas que no mantienen una única estructura estable en el tiempo, sino que presentan un conjunto de estructuras o ensamble conformacional, sin estructura terciaria y con bajo contenido de estructura secundaria [1]. Estas regiones, conocidas como regiones intrínsecamente desordenadas (IDRs, por sus siglas en inglés ***I**ntrinsically **D**isordered **R**egions*), juegan un rol central en procesos celulares [1]. Dos elementos funcionales presentes en las IDRs son los dominios intrínsecamente desordenados (IDDs, por ***I**ntrinsically **D**isordered **D**omains*) y los motivos lineales de secuencia corta (SLiMs, por ***S**hort **Li**near **M**otifs*) que se describirán en detalle más adelante [2] .

Los dominios globulares, IDDs y SLiMs son tres módulos que median interacciones proteína-proteína y presentan características funcionales y estructurales específicas [2]. Si bien existen excepciones, los dominios globulares pueden establecer interacciones de alta afinidad con otros dominios globulares (Figura 1.1A), que pueden estar implicadas por ejemplo en la estructura o andamiaje celular [2]. En contraste, los IDDs y SLiMs suelen establecer interacciones débiles o transientes con dominios globulares (Figura 1.1B, C) típicamente en el rango micromolar y desempeñan, por ejemplo, funciones de señalización o regulación que requieren la formación y ruptura de las interacciones [2]. En relación a las características estructurales, los dominios globulares presentan una conformación estable en estado nativo mientras que los IDDs y SLiMs sólo adquieren una conformación estable al unirse con el dominio globular correspondiente [1].

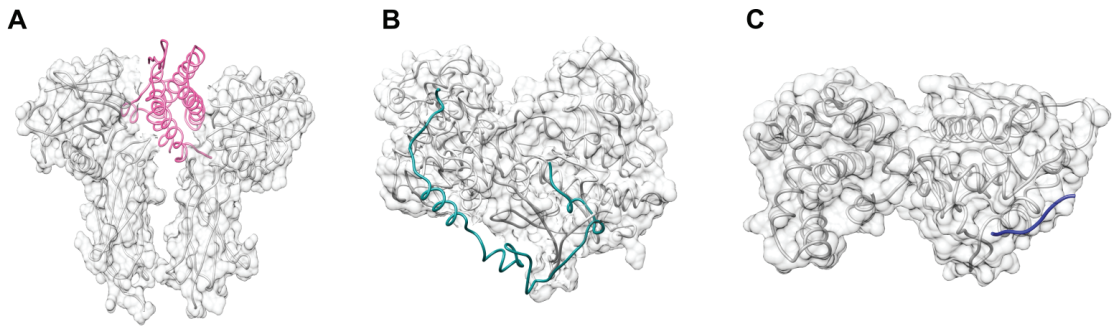


Figura 1.1. Ejemplos representativos de interacciones proteína-proteína entre distintos módulos de interacción. A: Interacción entre dominios globulares: complejo de interacción de eritropoyetina (rosa) con su receptor de membrana (gris) (PDB: 1CN4) [3]. **B:** Interacción entre dominio globular y dominio intrínsecamente desordenado (IDD): complejo de interacción del inhibidor 1B de CDK (gris) con el segmento desordenado del complejo Ciclina-A2/CDK2 (cyan) (PDB: 1JSU) [4]. **C:** Interacción entre dominio globular y SLiM: complejo de interacción del dominio *pocket* de retinoblastoma (gris) con el SLiM LxCxE presente en la proteína E7 de HPV (azul) (PDB: 1GUX) [5].

En resumen, existen diferencias entre los dominios globulares y los IDDs y SLiMs que reflejan la función dinámica que desempeñan los IDDs y SLiMs en los procesos celulares. El presente trabajo se enfoca en particular en las interacciones entre un dominio globular y dos SLiMs.

1.2. Proteínas intrínsecamente desordenadas

Siguiendo el paradigma clásico de secuencia-estructura-función, las funciones de las proteínas fueron atribuidas durante muchos años a los dominios globulares ya que adquirían una estructura tridimensional que varía poco alrededor de un estado de equilibrio. Actualmente, la noción de que las regiones desordenadas al carecer de estructura, carecen de función es constantemente desafiada por la evidencia acumulada de funciones claves desempeñadas por IDRs [6]. Aproximadamente un tercio del proteoma humano está compuesto por IDRs, pero aún la mayoría de estas regiones no están caracterizadas [1,2]. La mayoría de las anotaciones de proteínas está basada en la información de familias de secuencias y dominios estructurados [1]. Por lo tanto, las diferencias entre las IDRs y las proteínas globulares hacen necesario el desarrollo de nuevas herramientas que permitan la anotación y predicción de funciones de IDRs.

Variabilidad de Secuencia en IDRs. A diferencia de las proteínas globulares que presentan una alta conservación de secuencia debido a que deben mantener una estructura terciaria, las IDRs se caracterizan por una mayor variabilidad de secuencia ya que carecen de restricciones estructurales permitiendo mayores tasas de inserciones y deleciones [1]. También existen diferencias a nivel de composición de secuencia. En las proteínas globulares prevalecen aminoácidos hidrofóbicos que facilitan la formación del núcleo de plegado, mientras que las IDRs se encuentran enriquecidas en aminoácidos polares o con carga que estabilizan la exposición al solvente, y en prolina que impide la formación de estructuras secundarias [1,7]. A diferencia de las proteínas globulares, las IDRs suelen

tener regiones de baja complejidad de secuencia, es decir, regiones con repeticiones de aminoácidos como dipéptidos y tripéptidos [1,8]. Estas diferencias y la baja conservación de secuencia observada en las IDRs debido a la falta de una restricción estructural, dificulta la asignación de funciones [1].

Características Estructurales. La mayoría de las proteínas en organismos eucariotas son modulares, combinando dominios globulares estructurados con regiones intrínsecamente desordenadas (IDRs) y sólo una minoría son completamente estructuradas o completamente desordenadas (IDPs o ***I**ntrinsically **D**isordered **P**roteins*) [1,2]. Los dominios globulares presentan una estructura terciaria rígida, compacta y plegada que se encuentra definida por la secuencia de aminoácidos que la componen y está relacionada con una función [9]. Las IDRs, por otro lado, carecen de una estructura terciaria definida y se caracterizan por presentar flexibilidad estructural [10]. Cuando interactúan con otras proteínas, las IDRs pueden adquirir una conformación ordenada [1]. En ausencia de interacción, las IDRs muestran un espectro de conformaciones desordenadas, caracterizadas por la adopción de diversas estructuras no plegadas [11]. Es posible identificar la estructura de un dominio a través de técnicas de cristalización con rayos X o mediante resonancia magnética nuclear (RMN) [12]. Las IDRs en las estructuras determinadas por rayos X no pueden ser asignadas y se anotan como *missing residues*. Por otro lado, las técnicas de identificación de estructuras por RMN brinda información sobre la dinámica y la estructura de una proteína en solución, permitiendo obtener múltiples conformaciones de una región IDR. Los dominios globulares muestran una conformación única en los modelos estructurales (Figura 1.2A), mientras que las proteínas con un grado de desorden intermedio pueden alinear algunos segmentos de sus modelos, aunque otros muestran múltiples conformaciones (Figura 1.2B). En las proteínas con un alto grado de desorden, prácticamente todos los modelos presentan diversas conformaciones debido a la flexibilidad de sus regiones (Figura 1.2C).

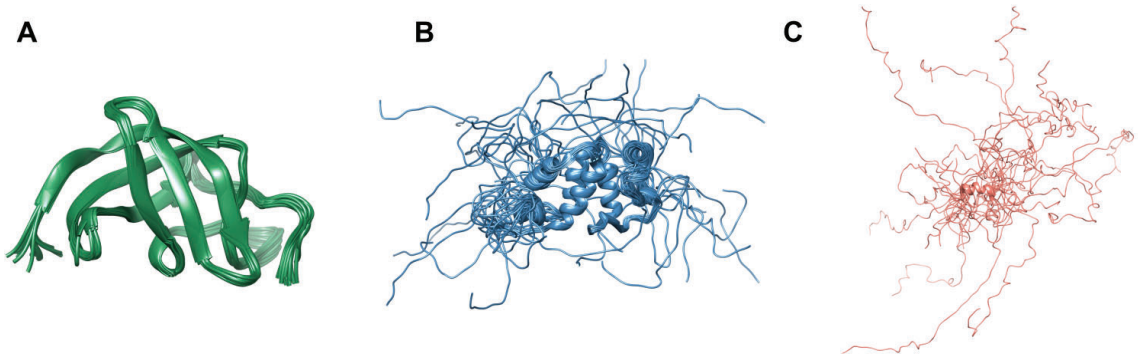


Figura 1.2. Ejemplos representativos de estructuras con distinto grado de orden resueltas por RMN. A: Segmento globular con una única conformación del dominio de unión al ADN de la proteína Y Box-1 (PDB: 6LMS) [13]. **B:** Segmento de desorden intermedio con regiones alineadas y otras de múltiples conformaciones del dominio BEN de la proteína humana NAC1 (PDB: 6LMS) [13]. **C:** Esquema de múltiples conformaciones que adopta la fosfoproteína soluble de tilacoide de espinaca, con un alto grado de desorden (PDB: 2FFT)[14].

Las IDPs pueden adoptar múltiples conformaciones que van a depender de las condiciones del ambiente, modificaciones post-traduccionales (PTMs, ***P**ost-**T**ranslational **M**odifications*), o la

interacción con un dominio globular [1]. Un ejemplo de cómo una PTM puede alterar la estructura y función de una proteína, es la proteína *pocket* Retinoblastoma (Rb). Rb tiene un dominio central llamado *pocket*, un dominio estructurado en el extremo N-terminal (RBN) y un IDD en el extremo C-terminal (RBC) (Figura 1.3A). Rb presenta IDRs en *loops* intra-dominios en los dominios RBN y *pocket,* y en *linkers* flexibles que conectan ambos dominios. Las secuencias *linker* contienen sitios de fosforilación por quinasas dependientes de ciclinas (CDKs) que son importantes para la inactivación de Rb. Al ser fosforilados estos sitios, provocan que los dominios RBN y RBC se plieguen sobre el dominio *pocket* (Figura 1.3B). Este cambio conformacional oculta los sitios de unión a otras proteínas, inhibiendo las interacciones con Rb [15].

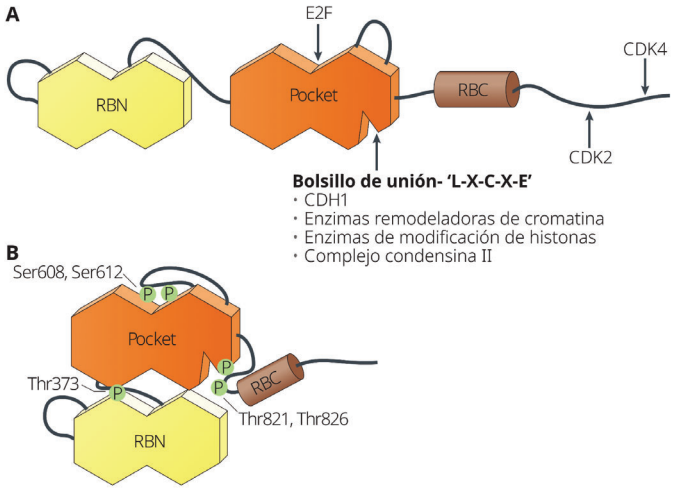


Figura 1.3. Alteración de estructura y función de Rb por PTMs. A: Diagrama de la estructura de dominios de la proteína Rb. Se esquematiza el dominio N-terminal (RBN), el dominio central *pocket* y el dominio C-terminal (RBC). Se señalan sitios de PTM por parte de quinasas dependientes de ciclinas 2 y 4 (CDK2, CDK4) y sitios de unión a los SLiMs E2F y LxCxE. A través del bolsillo de unión LxCxE Rb interactúa con proteínas como la proteína homóloga CDC20 1 (CDH1), enzimas remodeladoras de cromatina y modificadoras de histonas y el complejo condensina II. **B:** Al ser fosforilada en los residuos Ser608, Ser612, Thr373, Thr821 y Thr826, el dominio *pocket* se pliega sobre el dominio RBN y el dominio RBC sobre el dominio *pocket*, enmascarando los sitios de unión de proteínas conteniendo a los SLiMs LxCxE y E2F. Figura adaptada de Dick-Rubin et al. 2013 [15].

Funciones de IDPs y Versatilidad Funcional. Las IDPs están implicadas en diversas funciones celulares, como la regulación de la transcripción y la traducción, la señalización y la organización de complejos multiproteicos [10]. El contexto desordenado de las IDPs les permite unirse a múltiples dominios proteicos, a menudo a través de varios SLiMs presentes en una misma IDR, que incluso pueden solaparse debido a su pequeño tamaño [1]. La presencia de más de un SLiM en una sola IDP les permite interactuar con distintos dominios, facilitando la formación de redes de señalización dinámicas [1].

Proteínas Desordenadas y Patologías. Aunque pocas IDPs se encuentran caracterizadas, el enriquecimiento de regiones desordenadas en organismos complejos sugiere que son de gran

importancia biológica [1,2] Muchas enfermedades neurodegenerativas se encuentran asociadas a la agregación de IDPs [16]. Entre las enfermedades más conocidas que causan demencia, se encuentran el Alzheimer y Parkinson. El Alzheimer, una enfermedad caracterizada por la pérdida de memoria progresiva, está relacionada con el agregado de 'placas seniles' en el espacio extracelular, compuestas por segmentos de la proteína precursora de amiloide y agregados de la proteína tau, ambas IDPs [11]. La enfermedad del Parkinson, es causada por el agregado de cuerpos de Lewy, principalmente compuestos por la IDP α-sinucleína [11]. Las IDRs también se encuentran asociadas a otras patologías como cáncer y patologías de origen viral. Por ejemplo, el virus del papiloma humano (HPV), causante del cáncer cervical, logra secuestrar la maquinaria celular por la interacción de SLiMs presentes en el IDD de la proteína E7 con numerosas proteínas celulares.

En resumen, las IDRs poseen funciones relevantes y se diferencian en composición de secuencia y características estructurales de las proteínas globulares, dificultando el uso de las herramientas bioinformáticas desarrolladas para la anotación automática de funciones en proteínas globulares. Muchas de las funciones biológicas de las IDRs están mediadas por SLiMs, que son el objeto de estudio en este trabajo y que se detallan en la siguiente sección.

1.3. Los motivos lineales o SLiMs

Los motivos lineales o SLiMs (por sus siglas en inglés ***S**hort **Li**near **M**otif**s***) son elementos modulares compuestos de una secuencia lineal y corta de entre tres y diez aminoácidos de longitud [17]. Los SLiMs se encuentran en IDRs y en segmentos flexibles de las proteínas o *loops* que conectan dominios globulares. Los SLiMs no poseen una estructura definida al estar libres, sino que adquieren estructura al interactuar con un dominio globular [2]. Por su localización dentro de una IDR los SLiMs están expuestos al solvente y accesibles para interactuar con otras proteínas [17]. A modo de ejemplo, en la Figura 1.1C, se muestra el dominio globular de la proteína retinoblastoma, representada como superficie en color gris, interactuando con el SLiM LxCxE, que se encuentra representado en color azul.

1.3.1. Propiedades y funciones de los SLiMs

Uno de los ejemplos más conocidos y tempranamente descritos de SLiMs son las señales o etiquetas de localización o *sorting* subcelular. Algunos ejemplos de estos SLiMs son la señal de localización y exportación nuclear (NLS, ***N**uclear **L**ocalization **S**ignal* y NES, ***N**uclear **E**xport **S**ignal*, respectivamente), cuya función es transportar moléculas hacia dentro y fuera del núcleo celular, y el SLiM KDEL, relacionado con el tráfico de proteínas desde el aparato de Golgi hacia el retículo endoplasmático (señal de retención en retículo) [2]. Además de las señales de localización subcelular, algunos SLiMs codifican para sitios de clivaje proteolítico o para PTMs como por ejemplo las

glicosilaciones, fosforilaciones y acetilaciones. Por último, los SLiMs también pueden codificar funciones de "docking" reclutando sustratos para su modificación enzimática, como por ejemplo el SLiM de *docking* presente en sustratos de la familia de quinasas MAP (MAPK, ***MAP K****inasa*) o de quinasas dependientes de ciclinas (CDKs) o funciones de "ligando" facilitando la formación de complejos macromoleculares al mediar la unión a dominios globulares de otras proteínas [1,2,18]. En esta última categoría de "ligando" se encuentran los SLiMs E2F y LxCxE, que permiten la formación de complejos macromoleculares con las proteínas *pocket* y son objeto de estudio de este trabajo.

Aminoácidos que Componen los SLiMs: La interacción entre un SLiM y un dominio globular está mediada por 5-10 aminoácidos que representan el centro o ***core*** del SLiM y establecen contacto directo con la superficie de un dominio globular. Dentro de la región ***core***, podemos distinguir a las ***posiciones fijas*** y las ***posiciones variables*** y en torno al ***core*** del SLiM podemos definir las ***posiciones flanqueantes*** (5-10 residuos al N- y C-terminal del SLiM). Estos tres tipos de posiciones en conjunto determinan la afinidad y especificidad de unión del SLiM por el dominio globular.

Las ***posiciones fijas*** son determinantes para la unión ya que establecen contactos directos con el dominio y, por lo tanto, sólo aceptan sustituciones conservativas, es decir, aminoácidos con características fisicoquímicas similares que poseen una alta complementariedad con la superficie de unión en el dominio globular [2]. En contraste, las ***posiciones variables*** del SLiM representan residuos que no interactúan con la superficie del dominio globular si no que apuntan hacia el solvente, y por lo tanto aceptan sustituciones no conservativas que pueden modular la afinidad de unión. Las ***posiciones flanqueantes*** establecen interacciones más variables y dinámicas con el dominio globular y también pueden modular la afinidad del SLiM por el dominio globular [2]. Por ejemplo, en la región flanqueante al SLiM PIP presente en polimerasas media la unión a la proteína "clamp" PCNA (por ***P****roliferating* ***C****ell* ***N****uclear* ***A****ntigen*) que recluta proteínas a la horquilla de replicación del ADN existen residuos con carga positiva que aumentan la afinidad de unión a PCNA por complementariedad con una superficie de PCNA enriquecida en residuos de carga negativa [19,20].

En resumen, los SLiMs median la formación de complejos proteicos y la regulación de procesos celulares mediante PTMs. Los SLiMs poseen unos pocos residuos que conforman el ***core*** del SLiM y determinan la interacción específica con un dominio globular o familia de dominios. El contexto de secuencia, es decir, las ***posiciones variables y flanqueantes*** pueden modular la afinidad. El tamaño acotado de un SLiM permite codificar múltiples funciones en una IDR de una proteína, por lo cual la identificación de SLiMs es fundamental para revelar nuevas funciones de las IDRs.

1.4. La familia de proteínas *pocket*

La familia de proteínas *pocket* está compuesta por tres proteínas parálogas: Retinoblastoma (Rb), p107 y p130 (Figura 1.4 A), que se encuentran involucradas en la regulación del ciclo celular y

la diferenciación celular [15]. La proteína *pocket* más estudiada es Rb, que fue identificada como una proteína supresora de tumores por su actividad en la regulación del ciclo celular y su rol central es controlar el checkpoint G1/S, restringiendo el inicio de la fase S de replicación del ADN [15,21].

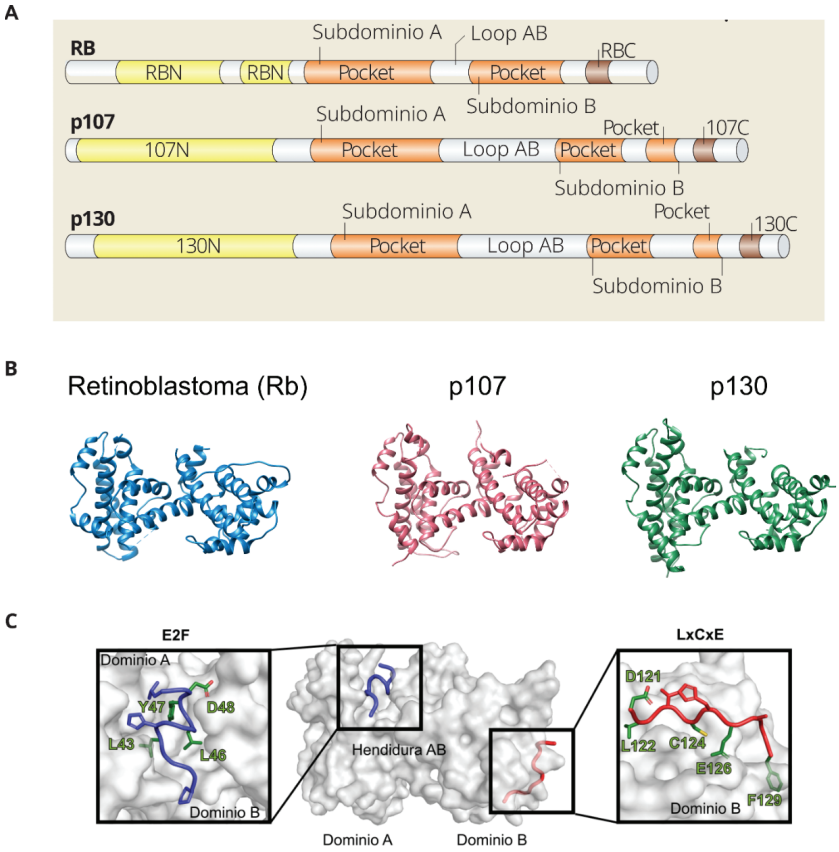


Figura 1.4. Familia de proteínas *pocket*. A: Las proteínas *pocket* Retinoblastoma (RB), p107 y p130 presentan similitud de dominios entre sí. Se observa un dominio N-terminal (RBN, 107N y 130N), el dominio *pocket* central conformado por los subdominios A y B conectados por el loop AB, y un dominio C-terminal (RBC, 107C y 130C) conectados por *loops* flexibles. Los números indican la posición de inicio y fin de cada segmento. Figura adaptada de Dick and Rubin 2013 [15]. **B:** Representación esquemática de dominios *pocket* altamente conservados entre Retinoblastoma (PDB 1GUX [5]), p107 (PDB 4YOZ [22]) y p130 (Modelo AF-Q08999-F1 obtenido de la base de datos de estructuras proteicas de AlphaFold [23]). El Loop AB está ausente en las estructuras. **C:** Estructura representativa del dominio *pocket* de Rb unido al SLiM E2F (azul) y LxCxE (rojo) de la proteína E1A de Adenovirus. Las cadenas laterales de los residuos principales del SLiM E2F y LxCxE están representadas como bastones. Figura de González-Foutel et al. 2022 [24].

La proteína Rb comparte aproximadamente un 25% de identidad de secuencia con p107 y p130, mientras que p107 y p130 comparten un 54% de identidad entre ellas [15]. Además de la alta similitud de secuencia, las proteínas *pocket* comparten una estructura de dominios que consiste en un dominio N-terminal (RBN), un dominio central *pocket* cuya estructura se encuentra conservada en las tres proteínas y le da el nombre a la familia (Figura 1.4B) y un dominio C-terminal (RBC) con múltiples sitios blancos de modificación post-traduccional (Figura 1.4A) [15].

El dominio *pocket* está formado por dos subdominios A y B , unidos por un *loop AB* flexible. La interfaz entre los subdominios, denominada hendidura AB está altamente conservada. En la

hendidura AB se une el SLiM E2F (Figura 1.4 C, izquierda en azul) presente en los factores de transcripción E2F [15]. El subdominio B presenta un bolsillo que interactúa con proteínas que contienen el SLiM LxCxE, como la histona deacetilasa 1 (HDAC1) y 2 (HDAC2), demetilasa 5-A lisina-específica (KDM5A) (Figura 1.4 C, derecha en rojo) [15]. **El grado de conservación del dominio *pocket* y de los bolsillos de unión de los SLiMs E2F y LxCxE entre las tres proteínas parálogas y entre especies sugiere que sus interacciones juegan un rol crucial en la regulación del ciclo celular** [15].

1.4.1. Funciones de las proteínas *pocket*

Las tres proteínas *pocket* comparten la función de regular el ciclo celular al inhibir la proliferación [15,21]. Esta actividad se logra en gran parte mediante su asociación con los factores de transcripción de la familia E2F, que controlan la expresión de genes implicados en la progresión del ciclo celular. En los tres casos, estas proteínas son inactivadas a través de la fosforilación mediada por quinasas dependientes de ciclinas (CDKs) [15,21].

La proteína Rb es la única proteína de la familia con actividad supresora de tumores e interactúa específicamente con miembros activadores de la familia E2F: E2F1, E2F2 y E2F3, y con los complejos de ciclinas D/CDK4 y D/CDK6 [21]. Las CDKs modulan la actividad de Rb por fosforilación a lo largo del ciclo celular, alterando su estructura, lo que provoca que Rb se disocie de los factores de transcripción E2F (Figura 1.5) [15]. Cuando Rb se encuentra hiperfosforilada, libera a E2F desencadenando la transcripción de genes dependientes de E2F, promoviendo la transición del ciclo celular de la fase G1 a la fase S [25]. Las proteínas p107 y p130, a diferencia de Rb, se asocian exclusivamente a los factores de transcripción E2F4 y E2F5 y a los complejos de ciclinas A/CDK2 y E/CDK2 [25,26]. Sus funciones se asocian principalmente a la diferenciación celular y la regulación del ciclo al interactuar, por ejemplo, con el complejo DREAM y las proteínas MuvB [22].

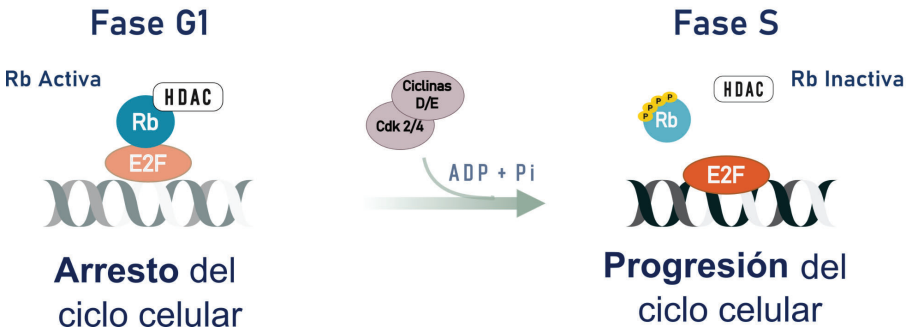


Figura 1.5. Regulación del ciclo celular por interacción de Rb con factores de transcripción E2F. En la fase G1 del ciclo celular Rb se encuentra en estado activo, unida a los factores de transcripción E2F a través de la hendidura AB y a HDAC a través del bolsillo de unión al SLiM LxCxE, provocando el arresto del ciclo celular. El avance del ciclo hacia la fase S provoca la fosforilación de Rb por parte de las CDKs 2 y 4 y su inactivación, liberando a E2F y HDAC se sus sitios de unión y promoviendo la activación de genes necesarios para la replicación celular. La desregulación de este mecanismo de acción provoca la proliferación de células tumorales. Figura adaptada de Dick and Rubin 2013 [15].

El rol más estudiado de Rb está relacionado con la regulación del ciclo celular de la fase G1 a S, sin embargo estudios recientes muestran que la inactivación de Rb produce fallas en la segregación cromosómica y la consecuente transformación maligna de células tumorales sugiriendo que Rb cumple roles no canónicos durante la mitosis [25,27]. Sin embargo, los mecanismos subyacentes son aún poco comprendidos. Otros estudios muestran que los niveles de p107 durante la fase S son críticos, sugiriendo que p107 también podría cumplir roles regulatorios cuyos mecanismos moleculares son desconocidos [28].

Estas funciones no canónicas de las proteínas pocket sugieren que aún hay numerosos interactores por identificar que explicarían los mecanismos moleculares de estas funciones.

1.4.2. Los SLiMs de unión a las proteínas *pocket*

Como se describió anteriormente, las tres proteínas *pocket* comparten un dominio central que media interacciones con los SLiMs LxCxE y E2F (Figura 1.6, A), presentes en los interactores.

Representación de los SLiMs: A nivel bioinformático, las características de secuencia de los SLiMs pueden representarse de diferentes maneras como por ejemplo, las **expresiones regulares** o los **logos de secuencia**. Las expresiones regulares pueden utilizarse para identificar SLiMs en archivos de secuencia rápidamente, mientras que los logos de secuencia permiten una visualización rápida de la variabilidad presente en un conjunto de secuencias que contienen un SLiM y permiten puntuar secuencias según su similitud al logo. Como ejemplo, presentaremos los SLiM de unión a las proteínas pocket.

Expresiones Regulares: Una expresión regular (o *regex*, por la contracción de las palabras en inglés ***reg****ular* ***ex****pression*) es un patrón de caracteres que representa una secuencia que puede ser identificada en una cadena de texto. Las *regex* fueron inicialmente incorporadas en el campo de la informática en los años ‘60 para la búsqueda de patrones en un editor de texto [29]. Desde entonces son ampliamente utilizadas en un gran número de aplicaciones informáticas y existen diversas sintaxis, aún cuando en 1992 fueron incorporadas a la familia de estándares POSIX (siglas que provienen de su nombre en inglés, ***P****ortable* ***O****perating* ***S****ystem* ***I****nterface for Uni****X***) definidos por la Sociedad Computacional IEEE para el mantenimiento de la compatibilidad entre diferentes sistemas operativos [30].

Las expresiones regulares de los SLiMs se construyen a partir de un alineamiento múltiple de secuencias (MSA) de instancias de un SLiM que fueron caracterizadas experimentalmente mediante análisis estructural (Figura 1.6, A), y mapeo por mutagénesis que permite determinar la región que comprende al SLiM y la influencia de diferentes residuos en la unión.

Como se explicó anteriormente, las ***posiciones fijas*** son determinantes para la unión y siguen un patrón definido de secuencia. En una expresión regular las ***posiciones fijas*** (en rojo en Figura 1.6,

B) se indican con el código de una letra del aminoácido permitido en esa posición o si se admite más de un aminoácido, se utilizan las letras correspondientes encerradas entre corchetes (Figura 1.6C) [31]. Las ***posiciones variables*** (en azul en Figura 1.6B) admiten mayor variabilidad y en la expresión regular están indicadas por un '.' (Figura 1.6C) [31].

Los alineamientos de secuencia evidencian el grado de conservación de las regiones flanqueantes al SLiM, hacia los extremos N y C-terminal del ***core***.

Logos de Secuencia: Otra forma de representar los SLiMs es mediante logos de secuencia (Figura 1.6D) los cuales se construyen a partir de un alineamiento de secuencias conteniendo el SLiM en estudio. En un logo de secuencia, el eje y representa el contenido de información de una posición (en bits), mientras que el eje x muestra la posición del alineamiento de secuencia. La altura de cada letra está relacionada directamente con la frecuencia relativa de ese residuo en esa posición y la altura total de cada posición está relacionada con el grado de conservación de los residuos en esa posición [32].

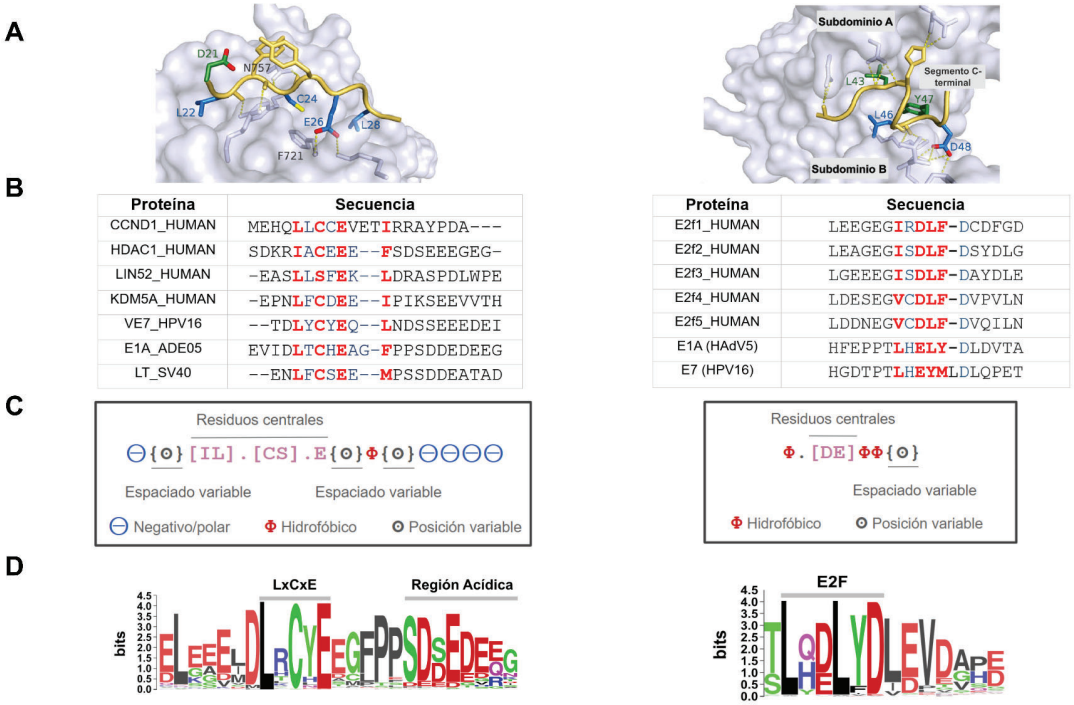


Figura 1.6. Representación de los SLiMs E2F y LxCxE. A: SLiM LxCxE de la proteína E7 de HPV16 (PDB: 1GUX [5]) (izquierda) y SLiM de la proteína E1A de Adenovirus (PDB: 2R7G [33]) (derecha), ambos unidos al dominio *pocket* de Rb representado como superficie gris. Se señalan en verde y azul las posiciones fijas. En amarillo, se muestran las posiciones variables y la segunda posicion fija en el caso del SLiM E2F. Figura adaptada de Palopoli 2018 [34]. **B:** Ejemplo de proteínas humanas y virales que poseen el SLiM LxCxE (izquierda) y E2F (derecha). En las secuencias se indican las posiciones fijas (rojo) y variables (azul) que definen la expresión regular. **C:** Representación esquemática de las expresiones regulares de los SLiMs LxCxE (izquierda) y E2F (derecha) donde se señala la región flanqueante al SLiM LxCxE que modula la interacción. Figura adaptada de Palopoli et al. 2018 [34]. **D:** Logos de secuencia de los SLiMs LxCxE (izquierda) y E2F (derecha). Figura adaptada de Glavina et al. 2018 [35].

El SLiM LxCxE

El SLiM LxCxE está presente en algunos interactores conocidos de las proteínas *pocket* como

las proteínas humanas histona desacetilasas (HDAC-1), ciclinas (ciclinas D1 y D2) y proteínas virales E7 de papilomavirus y E1A de adenovirus (Figura 1.6A, izquierda).

Los SLiMs de proteínas eucariotas se encuentran anotados y validados experimentalmente en la base de datos de Motivos Lineales Eucariotas (ELM, por sus siglas en inglés ***E**ukaryotic **L**inear **M**otif*) [36], que reúne la colección más grande de SLiMs, que actualmente alcanza las 4272 instancias clasificadas por tipo de SLiM, función y clases ELM, que se describen a través de expresiones regulares. La expresión regular del SLiM LxCxE está definida en ELM [36] como:

[DEST].{0,4}[LI].C.E.{1,4}[FLMIVAWPHY].{0,8}[DEST]

El centro o ***core*** de la expresión regular (en rojo) se encuentra definida por tres **posiciones fijas** y dos **posiciones variables** (simbolizadas por un punto). Las **posiciones fijas del core** admiten [IL], C y E son los determinantes de la unión (en rojo, Figura 1.6B, izquierda), dado que se entierran en el bolsillo de unión LxCxE, mediando la interacción entre la proteína que tiene el SLiM y el dominio *pocket* [37]. Las **posiciones variables** no entran en contacto con la superficie, siendo de preferencia en este SLiM los residuos aromáticos [FWY] por formar estructuras apiladas que se orientan hacia afuera del complejo y estabilizan la interacción (en azul, Figura 1.6, B izquierda). Existen variantes del SLiM LxCxE, por ejemplo, la proteína LIN52 que posee un SLiM LxSxE donde la cisteína central está reemplazada por una serina (S) [22].

Por fuera del *core* del SLiM se identifican posiciones **flanqueantes** que modulan la afinidad de la interacción con el dominio pocket. En el extremo N-terminal, se admiten residuos ácidos y polares que se orientan hacia afuera de la estructura, mientras que en el extremo C-terminal se admiten residuos hidrofóbicos que se entierran en el dominio *pocket* [37] (Figura 1.6, C izquierda). Estas posiciones pueden presentar conservación de residuos que se visualizan en los logos de secuencia (Figura 1.6, D izquierda).

El SLiM E2F

El SLiM E2F se encuentra presente en el dominio de transactivación de la familia de factores de transcripción E2F y en la proteína viral E1A de adenovirus (Figura 1.6, A derecha).

La expresión regular definida en ELM [36] para este SLiM es:

..[LIMVA].[DE][LMF][FYM][IL]{0,1}([DE]|(S)).

Donde el *core* del SLiM está representado por cinco residuos (en rojo). La primera **posición fija** admite residuos hidrofóbicos. La segunda **posición es variable** (en azul, Figura 1.6, B derecha) y le siguen tres **posiciones fijas** que admiten un residuo ácido y dos residuos hidrofóbicos alifáticos o aromáticos (en rojo, Figura 1.6, B,C derecha). El SLiM E2F forma una hélice anfipática pequeña en la que la primera, tercera y cuarta **posiciones fijas** quedan en una misma cara y se entierran en el bolsillo E2F del dominio *pocket*, mientras que la **posición variable** y la segunda **posición fija** se orientan en

otra cara de la hélice, quedando expuestos al solvente (Figura 1.6, A derecha) [34,38].

Otras regiones **flanqueantes** al *core* del SLiM (Figura 1.6, D derecha) probablemente modulen la afinidad de interacción con el dominio *pocket*, aunque a diferencia del SLiM LxCxE, los datos experimentales para el SLiM E2F son escasos hasta la fecha [34].

1.4.3. Limitaciones en la identificación de nuevas interacciones mediadas por SLiMs

Identificación mediante métodos bioinformáticos. Los SLiMs cumplen una amplia variedad de funciones moleculares y se estima que son muy abundantes, por ejemplo estimaciones actuales sugieren que existen más de 100.000 SLiMs funcionales en el proteoma humano de los cuales sólo el 5% es conocido [39]. Dado que los SLiMs median interacciones a través de un número limitado de residuos, computacionalmente es posible encontrar el mismo patrón de secuencia utilizando una expresión regular de dos o tres posiciones sin que ésto garantice que el SLiM es funcional. Es por esto que es necesario definir herramientas adicionales de filtrado tales como la localización celular o la exposición del SLiM candidato para interactuar con un dominio, para disminuir la aparición de falsos positivos y mejorar las chances de identificar SLiMs funcionales. La mejora de los métodos computacionales está limitada por la baja disponibilidad de datos experimentales sobre SLiMs que permitan el entrenamiento y desarrollo de mejores algoritmos bioinformáticos para detectarlos.

Identificación mediante métodos experimentales. Identificar SLiMs mediante técnicas experimentales convencionales, como el *pull down* o la proteómica presenta grandes limitaciones debido a que estos módulos pequeños suelen establecer interacciones débiles o transitorias que pueden perderse durante las técnicas de preparación de muestra. Además, las interacciones proteína-proteína mediadas por SLiMs usualmente requieren la presencia de otros módulos de interacción en las proteínas involucradas lo cual dificulta su identificación experimental [40,41].

Por lo tanto, podemos concluir que actualmente existen importantes limitaciones para descubrir nuevos SLiMs utilizando las herramientas computacionales y experimentales existentes.

1.5. Interactores de proteínas *pocket* y antecedentes del grupo de trabajo

Actualmente, se identificaron más de cien proteínas humanas capaces de interactuar con las proteínas *pocket* en la base de datos IntAct [42]. El gran número de blancos proteicos de las proteínas *pocket* y el hecho de que existen pocos mecanismos conocidos de interacción con esta familia de proteínas, sugiere que los SLiMs E2F y LxCxE están presentes en proteínas aún no identificadas, que

a su vez podrían explicar los roles no canónicos de Rb y p107. Muchas de las interacciones mediadas por SLiMs, depositadas en la base de datos de motivos lineales ELM [36], fueron identificadas experimentalmente mediante técnicas de pequeña escala (*low-throughput*), ya que la baja afinidad de las interacciones mediadas por SLiMs, dificultan la detección de estas interacciones en estudios de gran escala [40].

Con el fin de identificar nuevos interactores de las proteínas *pocket,* el grupo de trabajo realizó el ensayo de **ProP-PD** cuyos resultados se analizan en este trabajo.

1.5.1. El ensayo ProP-PD para identificar SLiMs a escala proteómica

La técnica de gran escala (*high throughput*) ProP-PD (***P**roteomic **P**eptide **P**hage **D**isplay*) es un ensayo de *phage display* modificado que permite identificar SLiMs de manera no sesgada y a escala proteómica utilizando la fusión de péptidos de 16 residuos procedentes del proteoma humano a proteínas de cápside de fago, utilizando tamaño de biblioteca de ~ 1.10^6 secuencias [43,44]. Este ensayo es una variante de la técnica de *Phage Display*, cuya modificación principal está en la construcción de la biblioteca de fagos (Figura 1.7). A diferencia del *phage display* clásico, la biblioteca no es aleatoria sino que incluye péptidos de 16 residuos que pertenecen a regiones desordenadas del proteoma humano (ver detalles en Sección 7.1). Los fagos que interactúan con los dominios *pocket* son seleccionados y secuenciados por la técnica de NGS para identificar los péptidos *hits*, para los cuales también se obtiene un número de cuentas. Los péptidos se encuentran diseñados mediante "tiling" o escalonado, de modo de superponerse entre sí, dando lugar a la ocurrencia de "overlaps" o regiones identificadas en más de un péptido. La técnica de ProP-PD es un método de punta que ha sido recientemente aplicado con éxito para descubrir nuevos SLiMs e interactores para Calcineurina [45], dominios PDZ [43,46], dominios globulares de SARS-CoV2 [47] y otros 30 dominios globulares más [40].

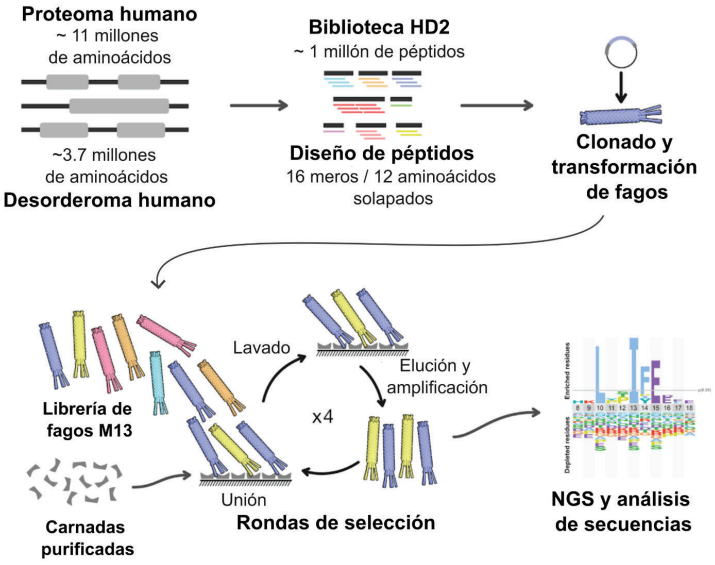


Figura 1.7. Esquema de la técnica de ProP-PD. Biblioteca de fagos de regiones desordenadas del proteoma humano expresados como péptidos de 16 aminoácidos, de los cuales 12 residuos se encuentran solapados con péptidos de una región contigua, logrando así cubrir con más de una copia los segmentos desordenados a analizar. Los péptidos se retro-transcribieron y añadieron adaptadores para construir una biblioteca de oligonucleótidos de fagos. La biblioteca de fagos fue sometida a un proceso de selección iterativo (rondas de selección), donde se la incubó con los dominios pocket de Rb, p107 y p130. Los péptidos enriquecidos durante la selección fueron secuenciados y analizados por NGS. Paneles adaptados de Benz et al. 2022 [40].

Con el fin de identificar nuevos interactores de las proteínas *pocket*, el laboratorio aplicó la técnica de ***ProP-PD*** en colaboración con los laboratorios del Dr. Norman Davey y la Dra. Ylva Ivarsson, desarrolladores de la misma (Figura 1.7). ***ProP-PD*** es una herramienta experimental poderosa para el descubrimiento de SLiMs candidatos [40]. En primer lugar, no está sesgada por sobreexpresión de proteínas, tipo o estadío celular y, en segundo lugar, permite la exposición de un dominio globular a la gran mayoría de las regiones desordenadas del proteoma humano.

En este ensayo, se expusieron los dominios *pocket* de Rb, p107 y p130 a una biblioteca de péptidos representando regiones desordenadas del proteoma humano que podrían contener un gran número de SLiMs por descubrir, incluyendo SLiMs candidatos de interacción con el dominio *pocket* [40]. Para el dominio *pocket* de Rb se obtuvieron un total de 308 péptidos de interacción o péptidos *hit*, para el de la proteína p107 un total de 103 péptidos y para p130 se obtuvieron 642 *hits*, sumando un total de 1053 *hits* para los dominios de la familia *pocket*.

Dado que ***ProP-PD*** es un ensayo *in vitro* entre un péptido fusionado a cápside de fago y un dominio globular, requiere experimentos adicionales para validar la interacción en el contexto de las proteínas enteras e *in vivo*. Sin embargo, el número de péptidos interactores es alto y es necesario realizar una priorización para su validación experimental, lo cual es el enfoque de este trabajo.

1.6. Hipótesis principal del trabajo

La familia de proteínas *pocket* está involucrada en procesos de regulación del ciclo celular, actuando como un punto de control en la progresión del cáncer a través de su interacción con proteínas que contienen SLiMs. Sin embargo, no existe un gran número de interactores reportados conteniendo SLiMs y se cree que aún quedan muchos por identificar [39]. El alto número de *hits* obtenidos en el ensayo ProP-PD sugiere la presencia de numerosos interactores con SLiMs funcionales aún no identificados, pero los métodos experimentales no permiten caracterizarlos a todos. De aquí se deriva la hipótesis principal del presente trabajo: **es posible identificar SLiMs funcionales capaces de interaccionar con las proteínas pocket Rb, p107 y p130 utilizando un ensayo de *Proteomic Peptide Phage Display (ProP-PD)* y priorizar los hits de este ensayo para su posterior validación experimental utilizando herramientas bioinformáticas.** Como resultado del análisis, podremos validar la calidad del ensayo ProP-PD y obtener una lista priorizada de SLiMs candidatos que interaccionan con las proteínas *pocket* que serán de utilidad para guiar las validaciones experimentales. En un futuro, estos análisis permitirán ampliar conocimientos acerca de la red de señalización de las proteínas *pocket*.

1.7. Objetivos

Objetivo general

El presente trabajo tiene como objetivo general la identificación y priorización de SLiMs E2F y LxCxE novedosos en el proteoma humano. Para alcanzar este objetivo se utilizarán métodos bioinformáticos existentes y de elaboración propia.

Objetivos específicos

OE1. Verificar la calidad del ensayo de ProP-PD aplicado a las *pocket* proteins.

OE2. Analizar la presencia de expresiones regulares definidas para los SLiMs E2F y LxCxE presentes en la lista de péptidos *hit* en el ensayo ProP-PD.

OE3. Evaluar características estructurales de las regiones donde se encontraron los *hits* que permitan diseñar una estrategia de filtrado.

OE4. Evaluar la estabilidad energética de péptidos *hit* utilizando matrices FoldX.

OE5. Definir un criterio de priorización de péptidos que fueron *hit* en el ensayo de ProP-PD utilizando las herramientas presentadas en los OE2, OE3 y OE4 con el fin de identificar a los mejores candidatos que serán ensayados experimentalmente en el laboratorio de trabajo.

Capítulo 2: Evaluación de la calidad del ensayo ProP-PD con *pocket* proteins

Como primer objetivo, se realizó una evaluación de la calidad del cribado realizado con la técnica ProP-PD, utilizando como carnada los dominios *pocket* de las proteínas Rb, p107 y p130. Una de las métricas de calidad establecidas en el ensayo ProP-PD es el número de cuentas obtenidos por NGS (***N***ext ***G***eneration ***S***equencing) para cada péptido, siendo de alta calidad los ensayos donde una gran proporción de *hits* tienen altas cuentas. Sin embargo, las cuentas de NGS no guardan relación directa con la afinidad de interacción [40] y algunos SLiMs validados, es decir, verdaderos positivos (TP, ***T****rue* ***P****ositive*) a veces dan bajas cuentas. En nuestro sistema, teniendo en cuenta el TP *hit* del ensayo con menor número de cuentas que es la instancia TP E2F3 con 6 cuentas registradas del análisis NGS, se calculó el porcentaje de péptidos con diez o más cuentas como criterio de corte. El 50% de los péptidos *hit* de Rb tienen diez o más cuentas. En el caso de p107, esto ocurre en el 62% de los péptidos mientras que en el de p130, sólo el 22% de sus *hits* cumple el criterio. Esto sugiere que el ensayo utilizado para p130 no es de alta calidad.

Para determinar la recuperación o *recall* de interactores conocidos se utilizaron tres estrategias: 1) se evaluó el *recall* de instancias conocidas de los SLiMs LxCxE y E2F reportados en la base de datos de referencia de SLiMs ELM [36,42]; 2) se evaluó el *recall* de interactores (proteínas enteras) reportados en la base de datos IntAct [36,42]; 3) se identificó cuántos de los interactores reportados en ensayos de gran escala (proteómica) realizados con la proteína Rb y no anotados en IntAct se encontraban representados entre los *hits* del ensayo [48]. En todos los casos, para calcular el *recall*, se tomó en cuenta el número de instancias de los interactores presentes en la biblioteca HD2 a partir de la cual se realizó el ensayo de Phage Display (ver Sección 7.1).

Brevemente, la biblioteca HD2 (siglas del inglés ***H****uman* ***D****esorderome* **2**) consiste en una biblioteca de péptidos de 16 residuos representando todas las regiones desordenadas (IDR) del proteoma humano. Para asegurar la cobertura y representación de todas las IDRs, los péptidos fueron diseñados con 12 residuos de solapamiento entre sí. Estos son expresados en la superficie de fagos que son incubados con los dominios *pocket* (ver Sección 7.1). Un procedimiento común en la construcción de bibliotecas de *phage display* es realizar la mutación de cisteínas por alaninas (C→A), evitando así la formación de puentes disulfuro en la superficie de los fagos. Dado que este trabajo se enfoca en la búsqueda de posibles interactores biológicos de las proteínas *pocket*, algunos de los cuales presentan en su secuencia al residuo C en lugar de A (variante de mayor afinidad), en este capítulo y en los siguientes los análisis se utilizarán las secuencias *wild-type* de los péptidos (variante conteniendo Cys).

2.1. Evaluación del *recall* de SLiMs conocidos (ELM)

Como resultado del ensayo de ProP-PD se obtuvieron 308 *hits* para la proteína Rb, 103 *hits* para p107 y 642 para p130. Estos resultados fueron analizados para evaluar la calidad y confianza del ensayo. Una forma de evaluar la calidad de ensayo consiste en analizar el número de SLiMs conocidos que se recuperan en el mismo. Con este objetivo, se realizó una búsqueda de SLiMs conocidos en la base de datos ELM, de donde se obtuvieron quince instancias conteniendo trece SLiMs LxCxE y dos SLiMs E2F de interacción con las proteínas *pocket* [36] (noviembre 2021). Estos datos se complementaron con tres instancias del SLiM E2F y una instancia del SLiM LxSxE recolectadas de literatura, alcanzando un total de 14 instancias del SLiM LxCxE y cinco instancias del SLiM E2F [34] (Tabla S1, Anexo). Si bien los datos se recolectaron en 2021, el número de SLiMs no ha cambiado hasta el presente y estas instancias se encuentran actualmente incorporadas a ELM [49].

Recall global de SLiMs para Rb y p107. Se desarrolló un programa para identificar cuáles de las instancias de los SLiMs LxCxE y E2F se encontraban presentes en la biblioteca HD2 y cuántas de estas instancias fueron recuperadas en el ensayo ProP-PD. De las 19 instancias, 17 se encuentran presentes en la biblioteca HD2 (Tabla S1, Anexo) y seis fueron recuperadas luego del ensayo (Tabla 2.1). Por lo tanto, se obtuvo un *recall* global del 35% (6/17) teniendo en cuenta el total de las instancias recuperadas utilizando los dominios *pocket* de Rb y p107 como carnada.

Tabla 2.1. Instancias recolectadas de ELM y literatura recuperadas en el ensayo ProP-PD.

Proteína	Secuencia[#]	SLiM*	HD2**	Hits***	
				Rb	p107
ARI4A_HUMAN	ET**LVCHE**VDLDDL	LxCxE			
CCND1_HUMAN	MEHQ**LLCCE**VETI	LxCxE			
CCND3_HUMAN	ME**LLCCE**GTRHAP	LxCxE			
EID1_HUMAN	TEE**LGCDE**IIDRE	LxCxE			
HDAC1_HUMAN	SDKR**IACEE**EFSD	LxCxE			
HDAC2_HUMAN	SDKR**IACDE**EFSD	LxCxE			
KDM5A_HUMAN	EPN**LFCDE**EIPIK	LxCxE		1	1
PPR26_HUMAN	SAE**LMCAE**AILDI	LxCxE			
PRDM2_HUMAN	E**IRCDE**KPEDLLE	LxCxE			
SMCA2_HUMAN	VER**LTCEE**EEEKI	LxCxE			
SMCA4_HUMAN	VER**LTCEE**EEEKM	LxCxE			
E2F1_HUMAN	EGEG**IRDLF**DCDF	E2F		1	
E2F2_HUMAN	AGEG**ISDLF**DSYD	E2F		3	1
E2F3_HUMAN	EEEG**ISDLF**DAYD	E2F		1	1
E2F4_HUMAN	ESEG**VCDLF**DVPV	E2F			
E2F5_HUMAN	DNEG**VCDLF**DVQI	E2F		1	
LIN52_HUMAN	LEAS**LLSFE**KLDR	Lx**S**xE			1
CCND2_HUMAN	ME**LLCHE**VDPVRR	LxCxE			
NDC80_HUMAN	YT**IKCYE**SFMSGA	LxCxE			

[#] Se subraya la secuencia que corresponde al SLiM.

* SLiM funcional reportado en ELM o en literatura [34].

** Una celda coloreada indica que la instancia está incluida en la biblioteca HD2 (N_{Total}: 17)

*** Una celda coloreada indica que el péptido fue *hit* de Rb o p107. En la celda se indica el número de péptidos solapados identificados como *hits* en el ensayo. Cada uno de estos *hits* se contó 1 sola vez independiente del número de péptidos recuperados.

Recall para cada proteína. Para Rb, no existe evidencia de interacción con la proteína Lin52 que contiene la variante LxSxE del SLiM LxCxE, por lo que no se considera en el total de instancias presentes en la biblioteca HD2 a la hora de realizar el cálculo de *recall*. Por lo tanto, se obtuvo un *recall* de 31% (5/16) para Rb (Figura 2.1A), de 24% (4/17) para p107 (Figura 2.1B) y de 6% (1/17) para p130 (no mostrado).

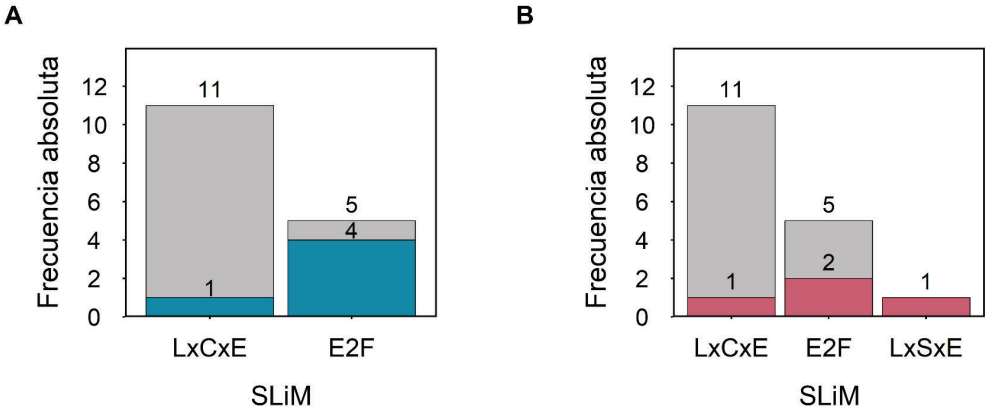


Figura 2.1. Interactores conocidos recuperados en el ensayo ProP-PD. Cantidad de péptidos recuperados utilizando Rb (**A**, en azul) y p107 (**B**, en rosa) como carnada conteniendo al SLiM LxCxE, E2F o LxSxE como caso exclusivo de p107. Las barras en gris indican la cantidad de péptidos reportados en bases de datos de interactores de las proteínas *pocket* que fueron incluidos en la biblioteca HD2.

Para Rb y p107 los valores de *recall* obtenidos en este trabajo, son más altos a los reportados en un *benchmarking* de la técnica ProP-PD, donde el *recall* promedio obtenido para más de 30 dominios proteicos es de 19,3% [43]. Este resultado indica que el cribado realizado para Rb y p107 es de muy buena calidad y que hay una alta probabilidad de identificar nuevos interactores dentro de los péptidos *hits*. Sin embargo, el bajo *recall* obtenido para p130 y el bajo número de cuentas que tienen la mayoría de los péptidos *hits* indica que el cribado usando p130 como carnada fue de mala calidad, por lo que no se continuó con el análisis de esta proteína.

Es importante resaltar que al contar ambas proteínas *pocket* para el SLiM LxCxE se recuperó el 17% (2/12) de instancias, mientras que para el SLiM E2F se recuperó el 100% (5/5) de las instancias. Este bajo *recall* para el SLiM LxCxE se debe a la mutación de cisteínas por alaninas (C→A) realizada durante la construcción de la biblioteca HD2. La mutación C→A en la segunda posición fija del SLiM Lx**C**xE da como resultado un SLiM Lx**A**xE y disminuye 60 veces la afinidad de unión al dominio *pocket* (Tabla S2, Anexo). La mayoría de los SLiM LxCxE del dataset de *benchmarking* tienen afinidades en el órden 10-50 μM [50] por lo que la mutación C→A bajaría la afinidad en casi dos órdenes de magnitud impidiendo su detección (Tabla S2, Anexo). Es por esto que la única instancia recuperada del SLiM LxCxE, es la proteína KDM5A que tiene afinidad nanomolar de unión a Rb y p107 [50]. Por lo tanto, concluimos que la mutación C→A realizada en HD2 genera un importante impedimento para el descubrimiento de SLiMs LxCxE.

2.2. Evaluación del *recall* de interactores conocidos (IntAct)

Como segunda forma de evaluar la calidad del ensayo, se realizó una búsqueda de interactores conocidos reportados en la base de datos IntAct, para los cuales se desconoce si poseen un SLiM LxCxE, un SLiM E2F o ningún SLiM conocido [42].

Para Rb se recolectaron 89 proteínas humanas (Tabla S3, Anexo) y para p107 se recolectaron 37 proteínas humanas (Tabla S4, Anexo) con al menos una interacción, directa o indirecta, reportada en IntAct [42] (Figura 2.2).

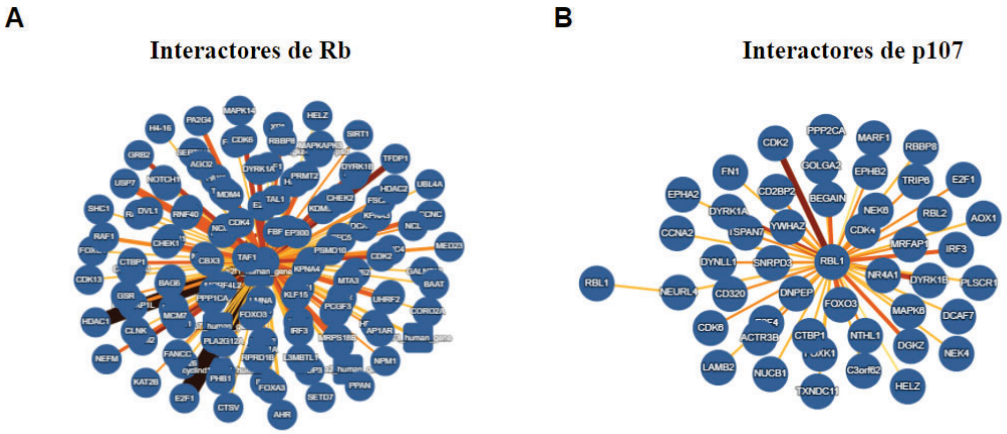


Figura 2.2. Red de interactores de las proteínas pocket en base de datos IntAct. Esquema de la red de interactores de las proteínas *pocket* reportados por la base de datos IntAct (noviembre 2021). **A:** Red de 89 interactores reportados para la proteína Rb. **B:** Red de 37 interactores reportados para la proteína p107.

Con el fin de identificar cuántos de estos interactores se encuentran presentes en la biblioteca HD2 y entre los péptidos *hit* de las proteínas *pocket*, se desarrolló un programa para detectar y anotar aquellos interactores con el mismo número de acceso. Dado que Rb es la proteína *pocket* más estudiada, existe un mayor número de interactores reportados que para p107. En base a esto y a que hay una alta conservación del dominio *pocket* entre las proteínas se tomaron las instancias E2F2, E2F4 y KDM5A reportadas para Rb como válidas para p107 también ya que son TP de p107. Para Rb, 87 de los 89 interactores se encuentran presentes en la biblioteca HD2 y ocho fueron recuperados luego del ensayo ProP-PD (Tabla 2.2). Para p107, 33 de los 37 interactores se encuentran presentes en la biblioteca HD2 y cinco fueron recuperados luego del ensayo ProP-PD (Tabla 2.2).

Si bien se conoce que la proteína LIN52 (LIN52_HUMAN) interacciona con p107 a través del SLiM LxSxE [22], hasta la fecha no se encuentra reportada en IntAct para esta proteína *pocket*, aunque sí lo está para p130. Por este motivo, será considerada como un interactor validado de p107 y se utilizará en el cálculo de *recall* para esta proteína *pocket*.

Tabla 2.2. Instancias reportadas en IntAct recuperadas en el ensayo ProP-PD.

	Proteína	Secuencia[#]	SLiM*	Interactores reportados en IntAct**	
				Rb	p107
Hits de Rb como carnada	ECD_HUMAN	SVMAPVDVDLNLVSNI	-	I	
	TAF1_HUMAN	SLITE**LTANE**ELTGTD	No reportado	I	
	XPA_HUMAN	LEVWGSQEALEEAKEV	-	I	
	KDM5A_HUMAN	EPN**LFCDE**EIPIKSEE	LxCxE	TP I	TP NR
	E2F1_HUMAN	DYHFGLEEGEG**IRDLF**	E2F	TP I+D	TP I
	E2F2_HUMAN	DYLWGLEAGEG**ISDLF**	E2F	TP I+D	TP NR
		GLEAGEG**ISDLF**DSYD	E2F		
		GEG**ISDLF**DSYDLGDL	E2F		
	E2F3_HUMAN	DYLLSLGEEEG**ISDLF**	E2F	TP I	TP NR
	E2F5_HUMAN	YNFNLDDNEG**VCDLFD**	E2F	TP I+D	
Hits de p107 como carnada	LIN52_HUMAN	TDLEAS**LLSFE**KLDRA	Lx**S**xE		TP NR
	E2F2_HUMAN	DYLWGLEAGEG**ISDLF**	E2F	TP I+D	TP NR
	E2F3_HUMAN	SD**LFDAY**DLEKLPLVE	E2F	TP I	TP NR
	E2F4_HUMAN	SELLEE**LMSSE**VFAPL	No reportado	I	I
		EE**LMSSE**VFAPLLRLS	No reportado		
	KDM5A_HUMAN	EPN**LFCDE**EIPIKSEE	LxCxE	TP I	TP NR

[#]Se destaca la secuencia del SLiM si está reportado en la base de datos ELM (negro) o bien donde se detecta la expresión regular (rojo).
* Clasificación del SLiM según ELM [36].
** La celda coloreada indica si la proteína es interactor de la proteína Rb (azul) o p107 (rosa). Se indica si es verdadero positivo (TP) y el tipo de evidencia reportada: I:Indirecta, D:Directa, NR:No reportado.

En base a estos resultados, concluimos que el *recall* de interactores para Rb es del 9 % (8/87 presentes en HD2) y para p107 es del 15 % (5/33). Estos valores son similares (Rb) o mayores (p107) al valor de 8,6 % obtenido en el benchmarking de ProP-PD usando IntAct como base de datos de referencia [43]. Este análisis provee de un mapeo candidato de la región de interacción para tres interactores de Rb reportados en IntAct (ECD, TAF1, XPA) e identifica un SLiM LxSxE novedoso en E2F4.

2.3. Evaluación del *recall* en ensayos de Proteómica

El grupo *Sanidas et al*. [48] realizó un ensayo de gran escala proteómico obteniendo un total de 432 interactores candidatos para Rb. El 88% (378/432) de estos candidatos se encuentran presentes en la biblioteca HD2 [48]. En el ensayo ProP-PD, se identificaron para Rb 14 péptidos *hits* que corresponden a un total de nueve proteínas reportadas en el ensayo de proteómica. De los 14 péptidos, tres son SLiMs de E2Fs ya reportados en la base de datos IntAct [42] y un total de 6 péptidos son SLiMs ya conocidos (Tabla 2.3).

Tabla 2.3. Interactores reportados en ensayos de proteómica.

ID Uniprot	Secuencia[#]	SLiM*	*Hit* de ProP-PD**	TP (ELM)**	Reportado en Intact**
E2F1_HUMAN	DYHFGLEEGEG**IRDLF**	E2F	■	■	■
E2F3_HUMAN	DYLLSLGEEEG**ISDLF**	E2F	■	■	■
E2F5_HUMAN	YNFNLDDNEG**VCDLF**D	E2F	■	■	■
E2F4_HUMAN	EEEG**ISDLF**DAYD	E2F		■	■
HDAC1_HUMAN	SDKR**IACEE**EFSD	LxCxE		■	■
HDAC2_HUMAN	SDKR**IACDE**EFSD	LxCxE		■	■
TPM3_HUMAN	**MMEAI**KKKMQMLKLDK	E2F	■		
DREB_HUMAN	S**LIDLW**PGNGEGASTL	E2F	■		
SCMC1_HUMAN	ELLKSYW**LDNFA**KDSV	E2F	■		
PRUN2_HUMAN	SG**IMELY**GSDIEPQPS	E2F	■		
	PT**FLEIW**NDSVDGDSF	E2F *like*	■		
	DRKTPT**FLEIW**NDSVD				
PDLI1_HUMAN	VTEEGKRHPYKMNLAS	-	■		
PLAK2_HUMAN	LTVVKDDDHGILDQFS	-	■		

[#] Si está reportado en la base de datos ELM (negro) o bien donde se detecta la expresión regular o similitud a ella (rojo).
* Tipo de SLiM identificado en la base de datos ELM [36] y si no están reportados a que SLiM corresponde;
** Una celda coloreada indica que el péptido fue *hit* en ProP-PD, o verdadero positivo reportado en ELM [36] o reportado en IntAct como interactor de la proteína Rb [42].

Además de los seis SLiMs conocidos, los otros ocho péptidos corresponden a seis proteínas del ensayo de proteómica (Tabla 2.3). Dos péptidos *hits* no poseen un SLiM conocido (proteínas PDLI1, PLAK2) y cinco poseen un SLiM E2F o similar al SLiM E2F (E2F-*like*) (proteínas TPM3, DREB, SCMC1, PRUN2) siendo SLiMs candidatos que aún no fueron validados. Los E2F-*like* no cumplen con la expresión regular de ELM, pero los aminoácidos comparten características fisicoquímicas similares a los del centro del SLiM.

2.4. Comparación de las técnicas analizadas

Cuando comparamos los resultados de todas las técnicas de gran escala en conjunto podemos analizar su grado de superposición. El número en común de interactores reportados para Rb en IntAct [42], ensayos de proteómica [48] y los obtenidos en el ensayo de ProP-PD (este trabajo) es bajo (Figura 2.3). Esta baja superposición parecería ser sorprendente, pero es común y ya fue descrita para otros sistemas estudiados por más de una técnica de gran escala [51]. Esta observación resalta las diferencias entre los ensayos e indica que ProP-PD es un ensayo complementario a los ensayos de proteómica y otras técnicas experimentales comúnmente utilizadas y que ProP-PD tiene el potencial de identificar nuevas interacciones mediadas por SLiMs no detectadas por las otras técnicas.

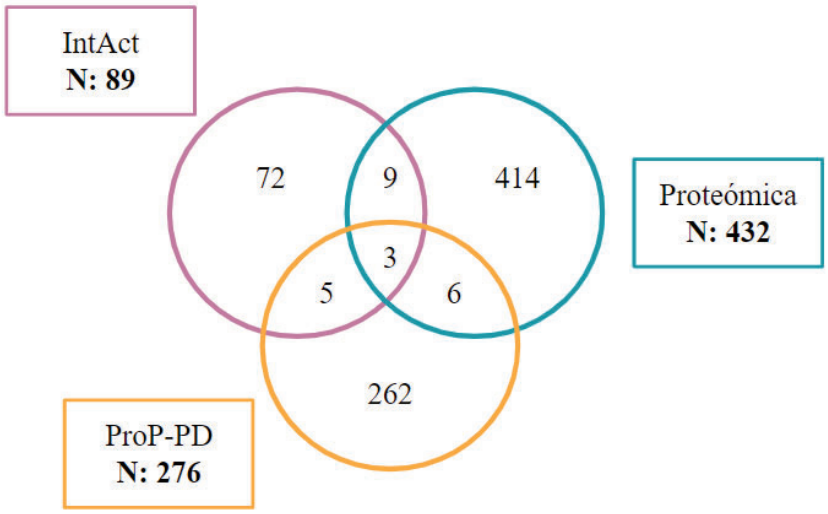


Figura 2.3. Comparación de técnicas de gran escala de interacciones proteína-proteína para la proteína Rb. Los números dentro de los recuadros (Intact, N=89; Proteómica, N = 432 y ProP-PD, N = 276) constituyen la cantidad de proteínas de los distintos set de datos. Las 262 proteínas que no se encuentran incluidas en otros set de datos y fueron *hit* del cribado, representan posibles interactores nuevos a identificar mediante técnicas experimentales a futuro.

2.5. Conclusiones sobre la calidad el ensayo de Prop-PD

El análisis de calidad realizado con los resultados del ensayo ProP-PD indica que el cribaje para Rb y p107 es de muy buena calidad, dado que el recall de SLiMs conocidos supera a lo reportado en el *benchmarking* de esta técnica, incluso considerando la baja recuperación de SLiMs LxCxE debido a la mutación C→A. Sin embargo, p130 presentó un bajo *recall* y la mayoría de sus péptidos presentan un bajo número de cuentas. Si bien la baja cantidad de instancias TP reportadas en ELM [36] para p130 dificulta el cálculo del *recall*, el bajo número de cuentas de los *hits* puede deberse al bajo grado de pureza de la muestra de p130 utilizada en el ensayo ProP-PD. Por este motivo, no se continuó con el análisis de p130 en las siguientes secciones del trabajo.

Para p107 y Rb, fue posible recuperar interactores validados en la base de datos IntAct [42] y encontrar coincidencias de interactores en ensayos de proteómica realizados para Rb [48]. Aunque no se espera una alta intersección de resultados entre técnicas de gran escala, este análisis muestra que fue posible complementarlas mapeando las regiones de unión y en algunos casos identificando SLiMs candidatos en interactores ya conocidos. Estos resultados avalan la alta calidad del ensayo, sugiriendo que muchos de los *hits* del ensayo ProP-PD podrían representar SLiMs novedosos presentes en interactores aún no conocidos de las proteínas *pocket*.

Capítulo 3: Detección de patrones de secuencia en péptidos *hits*

En este capítulo, se presenta un análisis utilizando variantes de los SLiMs LxCxE y E2F para identificar patrones de secuencia presentes en los péptidos *hit* de las proteínas Rb y p107. Estos patrones de secuencia determinan la afinidad de unión con los dominios *pocket*. Por lo tanto, este análisis nos permitirá priorizar péptidos conteniendo patrones más favorables, para ser testeados experimentalmente.

3.1. Enriquecimiento de SLiMs en los Péptidos *hit* utilizando MEME

Para analizar si existe una sobrerrepresentación de algún patrón de secuencia en los péptidos *hit,* se construyeron logos de secuencias con el total de *hits* para ambas proteínas *pocket* (Figura 3.1) utilizando MEME [52]. MEME identifica patrones repetitivos de longitud fija, a partir de un conjunto de secuencias.

En primer lugar, se analizó el total de los péptidos *hits* del ensayo de Prop-PD. Para esto, a partir de un archivo FASTA con las secuencias de los péptidos de 16 residuos se buscó en MEME patrones de 5 residuos de largo mínimo y 16 de largo máximo sin modificar los parámetros por defecto (Figura 3.1).

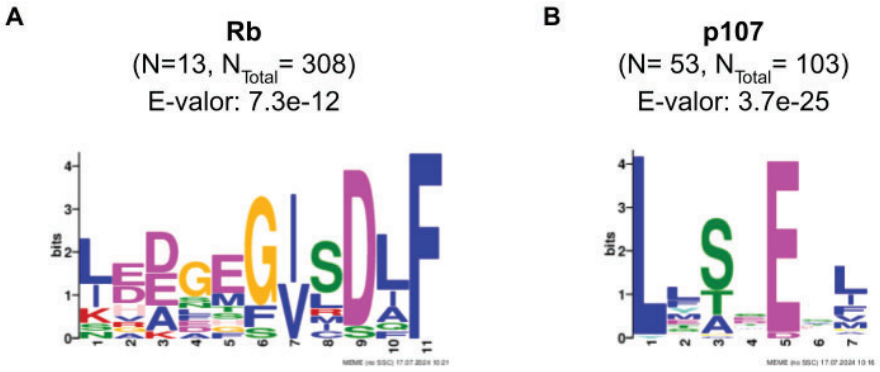


Figura 3.1. Logos de secuencia de péptidos *hit* de la proteínas *pocket*. Representación de residuos por posición presentes en secuencias de péptidos que fueron *hit* de las proteínas *pocket* en el ensayo ProP-PD utilizando la herramienta MEME [52]. El tamaño de cada letra representa la frecuencia de aparición del residuo en la posición indicada en el eje de abscisas. Los colores de cada símbolo corresponden al esquema de colores que MEME utiliza de acuerdo a similitudes bioquímicas de los aminoácidos. **A:** Logo de secuencias de 13 de 308 péptidos *hit* de Rb. **B:** Logo de secuencias de 53 de 103 péptidos *hit* de p107.

El alineamiento de secuencias que realizó MEME para construir el logo de los 308 péptidos

que fueron *hit* de Rb, incluyó el 4% (13 de 308 péptidos) de las secuencias (Figura 3.1A) mientras que en el caso de los 103 péptidos *hit* de p107, incluyó el 52% (53 péptidos) (Figura 3.1B). En este trabajo, se consideró 0,005 como máximo valor que puede adoptar el E-valor para que el logo sea significativo. Para el caso de Rb, se obtuvo un logo de E-valor de 7,3e-12, y para p107 el E-valor fue de 3,7e-25, y por lo tanto ambos son significativos. El bajo número de secuencias alineadas de péptidos *hit* de Rb se debe a que las secuencias con el SLiM E2F en el conjunto de datos, son probablemente más divergentes. Esto dificulta que MEME las alinee con secuencias altamente conservadas.

Este último resultado refleja la sobrerrepresentación del SLiM LxCxE en más de la mitad de los péptidos que fueron *hit* de p107. El patrón sobrerrepresentado posee un enriquecimiento de S en la segunda posición fija del SLiM Lx**C**xE. También se observa un enriquecimiento en residuos hidrofóbicos en la posición +2 con respecto al centro del SLiM, que según los ensayos experimentales mejora la afinidad de interacción con el dominio *pocket* [50,53]. Por el contrario, en el total de péptidos *hit* de Rb se observa la sobrerrepresentación del SLiM E2F en las posiciones siete a once del logo (Figura 3.1A) y no del SLiM LxCxE. La no detección del SLiM LxCxE, sugiere que hay un bajo número de péptidos con este SLiM.

El análisis de los péptidos que fueron *hit* de las proteínas *pocket* muestran sobrerrepresentación de los SLiMs E2F y LxCxE. En particular, se observó que la variante LxSxE se encuentra enriquecida al utilizar el dominio *pocket* de p107 como carnada y no el de Rb por ser ésta una interacción de menor afinidad (Tabla S2, Anexo). Estos resultados sugieren que los hits de ProP-PD representan los patrones conocidos de unión a proteínas *pocket*.

3.2 Detección de SLiMs mediante expresiones regulares

Para poder priorizar los péptidos *hit* conteniendo SLiMs conocidos de unión a proteínas *pocket* se desarrolló un programa de detección de expresiones regulares de los SLiMs LxCxE y E2F y se lo aplicó a los péptidos *hits* de ProP-PD.

Con el fin de capturar la mayor cantidad de péptidos *hits* con SLiMs, las expresiones regulares utilizadas son menos restrictivas que las definidas en la base de datos ELM [36] (ver Sección 1.4.2) dado que no incluyen las posiciones flanqueantes al ***core*** del SLiM y las posiciones fijas del ***core*** fueron degeneradas, es decir, presentan mayor cantidad de aminoácidos permitidos.

La expresión regular utilizada del SLiM LxCxE fue la siguiente:

`[IL].[CAST].E`

Para el SLiM E2F, la expresión regular utilizada fue:

`[IVLMA].[NQDE][IVLMAFYW][IVLMAFYW]`

En un análisis general de los 308 péptidos *hit* utilizando Rb como carnada, se detectaron 135 (44%) con al menos una expresión regular: en el 30% (91 péptidos) se detectó la expresión regular del SLiM E2F, en el 10% (32 péptidos) variantes del SLiM LxCxE y en el 4% (12 péptidos) ambos SLiMs. En el 56% (173 péptidos) de los péptidos *hit* no se detectó ningún SLiM con alguna de las expresiones regulares (Tabla 3.1). El bajo porcentaje de péptidos conteniendo variantes de la expresión LxCxE podría deberse a la mutación de la C de la segunda posición fija del SLiM por A durante la construcción de la biblioteca HD2, que disminuye la afinidad de interacción (Tabla S2, Anexo). Este resultado también explica por qué el patrón LxCxE no fue hallado por MEME para Rb.

Para p107, de los 103 péptidos que fueron *hit*, 63 (61%) cumplen con al menos una de las expresiones regulares definidas para los SLiMs: en el 8% (ocho péptidos) se detectó al SLiM E2F, un 44% (45 péptidos) presentan variantes del SLiM LxCxE, en el 10% (diez péptidos) se detectan ambos y un 39% (40 péptidos) no presenta expresiones detectables (Tabla 3.1). Este resultado también explica por qué el patrón E2F no fue hallado por MEME para p107.

Tabla 3.1. Detección de expresiones regulares en péptidos hit de las proteínas pocket.

	Rb* (N_T=308)	p107* (N_T=103)
SLiM	**N péptidos**	**N péptidos**
E2F	91	8
LxCxE	32	45
LxCxE y E2F	12	10
Sin SLiM	173	40

* Número de péptidos *hit* que presentan variantes de los SLiMs E2F, LxCxE o ambos, o sin SLiM detectado para Rb (N_{Total}: 308) y p107 (N_{Total}:103).

Los péptidos en los que se detectaron ambos SLiMs fueron incluidos tanto en el análisis de LxCxE como en el de E2F, dado que se desconoce cuál de los dos segmentos media la interacción con Rb o p107 en el ensayo ProP-PD. Por lo tanto, el total de *hits* con SLiM LxCxE detectados para Rb es de 44 (14%) y para p107 es 55 (53%), mientras que el total de péptidos con SLiM E2F detectado para Rb es de 103 (33%) y para p107 es 18 (18%).

La detección de SLiMs en los péptidos *hit* sugiere que la unión de muchos *hits* se encuentra mediada por SLiMs LxCxE y E2F y que el cribado realizado en el ensayo ProP-PD permite identificar interactores novedosos, que aún no han sido reportados.

3.2.1. Análisis de la variabilidad de secuencia del SLiM LxCxE presente en los péptidos *hits*.

En la sección anterior se detectó la secuencia del ***core*** del SLiM LxCxE en 44 péptidos *hits* de Rb y 55 péptidos *hits* de p107. A continuación, se analiza en primer lugar la variabilidad de secuencia observada en la segunda posición fija del ***core*** del SLiM y en segundo lugar la variabilidad de

secuencia observada en las posiciones variables y flanqueantes del SLiM LxCxE.

Análisis de variabilidad de secuencia en la segunda posición fija del core del LxCxE: Las posiciones fijas del core del SLiM son los puntos de contacto en la interacción entre el péptido y el bolsillo de unión al SLiM LxCxE del dominio *pocket*.

La primera (`[IL]`) y última posición fija (`E`) incluye residuos observados en proteínas celulares y virales con un SLiM LxCxE funcional, como por ejemplo, HDAC1 (IxCxE), ARID4A (LxCxE) y E7 de HPV (LxCxE). Por el contrario, la segunda posición fija (`[CAST]`) además de incluir la variante con serina observada en LIN52 (LxSxE), un interactor de p107, se incluyó el residuo A, ya que mientras C→S disminuye 220 veces la afinidad por Rb, la mutación C→A la disminuye 62 veces (Tabla S2, Anexo), indicando que LxAxE podría ser una variante aún no caracterizada del SLiM. Por último, se incluyó también la T en esta posición por su similitud estructural con S aunque aún no se haya reportado evidencia funcional de esta variante del LxCxE, y porque se halló este residuo enriquecido en los péptidos *hit* obtenidos para p107 (Figura 3.1).

Para evaluar si existe una diferencia en la preferencia de variantes de la segunda posición fija del SLiM LxCxE para Rb y p107, se tomaron los 44 *hits* de Rb (Figura 3.2 A) y los 55 *hits* de p107 (Figura 3.2 B) con un LxCxE detectado por expresiones regulares, buscando un patrón sobrerrepresentado con MEME.

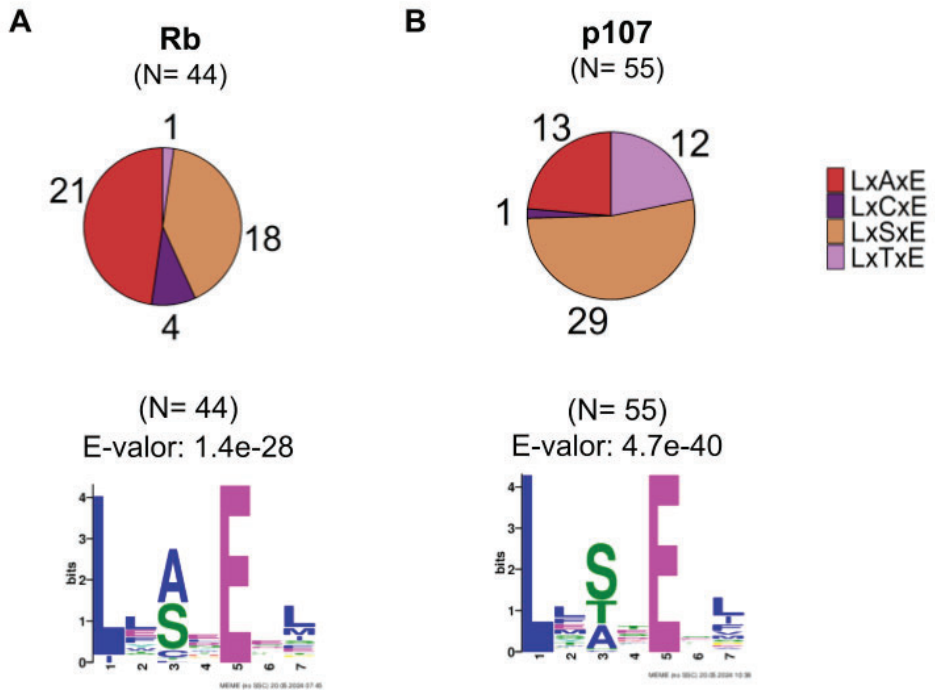


Figura 3.2. Análisis de expresiones regulares de péptidos *hit* con SLiM LxCxE detectado. Distribución de variantes [CAST] en la segunda posición fija del SLiM LxCxE (*arriba*) y logos de secuencia construidos en el sitio web MEME [52] (*abajo*) utilizando como carnada **A:** Rb y **B:** p107. Para Rb se analizaron 44 secuencias de péptidos *hit* y para p107 55 secuencias de péptidos *hit* con SLiM LxCxE detectado. En ambos casos para el gráfico de torta (*arriba*) por fuera se detalla el número de péptidos en los que se detectó cada variante del SLiM LxCxE y el número total de péptidos con el SLiM LxCxE detectado, y para MEME (*abajo*), se detalla el número de secuencias alineadas (N) y el E-valor del alineamiento. El tamaño de cada letra es directamente proporcional a la frecuencia de aparición del residuo en la posición indicada en el eje x. Los colores de cada letra son de acuerdo a características fisicoquímicas de los aminoácidos.

El análisis de la segunda posición fija del SLiM, mostró que para Rb las variantes LxSxE y LxTxE se encuentran representadas en un 43% (19 de 44) (Figura 3.2 A, arriba). Por el contrario, para p107 se observó un enriquecimiento del 75% (41 de 55) de estas variantes (Figura 3.2 B, arriba). La misma proporción de variantes se puede observar en los logos de secuencia realizados en el sitio web MEME [52] que muestran a la A con mayor frecuencia en la segunda posición fija del SLiM para Rb (Figura 3.2 A, abajo) y a S para p107 (Figura 3.2 B, abajo). Estos resultados son consistentes con estudios que reportan la preferencia de unión de p107 por el SLiM LxSxE [22,34].

Análisis de la variabilidad de secuencia en posiciones variables y flanqueantes

A partir de la expresión regular para el SLiM LxCxE definida en ELM [36] y junto con el conocimiento existente sobre los determinantes de afinidad de este SLiM, se definió un conjunto de expresiones regulares que amplían el ***core*** del SLiM LxCxE (ver Sección 7.4.1) [36] y corresponden a las características de mayor afinidad:

[DE][IL]**[YFH]**[CAST]**[YFH]**E.**{1,2}[WFILYVM]**

La afinidad del SLiM LxCxE es modulada por las posiciones variables del ***core*** del SLiM y los residuos flanqueantes al SLiM:

Residuo flanqueante ácido N-terminal: Se observó que péptidos virales con residuos ácidos en la posición inmediatamente anterior al *core* del SLiM (posición -1), se unen al dominio *pocket* con mayor afinidad que péptidos con cargas positivas en esa posición [50].

Posiciones variables en el *core* del SLiM LxCxE: [34,53]. Las posiciones variables ocupan la segunda y cuarta posición del ***core*** del SLiM LxCxE. Estos residuos se orientan hacia afuera de la superficie de contacto [22,34], formando una estructura apilada (*stacking*) e/o interaccionan entre sí, previniendo interacciones entre el péptido y el solvente, que desestabilizan la unión con el dominio *pocket [50]*. Los residuos de preferencia en las posiciones variables del SLiM LxCxE son [YFH].

Residuo hidrofóbico C-terminal: Residuos hidrofóbicos y aromáticos en las posiciones +2 o +3 del SLiM favorecen la alta afinidad de unión [50,53], en comparación a péptidos conteniendo estos residuos en las posiciones +4 o +5 [50].

En base a estas observaciones, se definieron cinco variantes del SLiM para su análisis, cuyo *ranking* debería reflejar SLiMs de mayor a menor afinidad (Tabla 3.2) (ver Tabla 7.2, Sección 7.4.1 y descripción del SLiM LxCxE en Sección 1.4.2). La **Variantes 5** es el core del SLiM. Las **Variantes 3, 4A, 4B** son expresiones regulares donde se consideran las características favorables de a una. La **Variante 1** combina las variantes 3, 4A y 4B, mientras que la **Variante 2** es la combinación de dos de las variantes. Las **Variantes 1 y 2**, representan variantes de alta afinidad del SLiM por el dominio *pocket* [22,34,50] (Figura 3.3A intersección entre diagramas de Venn).

Tabla 3.2. Definición de variantes del SLiM L.C.E.

Variante	Características[#]	Expresión regular
1	Res. ácido en -1, posición/es variable/s, res. hidrofóbico +2/+3	**[DE]**[IL]**[YFH]**[CAST].E.**{1,2}[WFILYVM]** **[DE]**[IL].[CAST]**[YFH]**E.**{1,2}[WFILYVM]**
2	**A.** Posición/es variable/s, res. hidrofóbico +2/+3	[IL]**[YFH]**[CAST].E.**{1,2}[WFILYVM]** [IL].[CAST]**[YFH]**E.**{1,2}[WFILYVM]**
	B. Res. ácido en -1, res. hidrofóbico +2/+3	**[DE]**[IL].[CAST].E.**{1,2}[WFILYVM]**
	C. Res. ácido en -1, posición/es variable/s	**[DE]**[IL]**[YFH]**[CAST].E **[DE]**[IL].[CAST]**[YFH]**E
3	Res. hidrofóbico +2/+3	[IL].[CAST].E.**{1,2}[WFILYVM]**
4	**A.** Res. ácido en -1	**[DE]**[IL].[CAST].E
	B. Posición/es variable/s	[IL]**[YFH]**[CAST].E [IL].[CAST]**[YFH]**E
5	Res. centrales o *core*	[IL].[CAST].E

Listado de las cinco categorías en orden decreciente de afinidad. Las variantes que incluyen más de una opción (2 A-C y 4 A y B), se asume que aumentan la afinidad del SLiM por el dominio *pocket* de manera similar.

De los 44 péptidos *hit* de Rb en los que se detectó el SLiM LxCxE se identificaron (Figura 3.3A):

- **Variantes 1 y 2**, en el 37% (16 péptidos) de los *hits*, con características que le confieren la

mayor afinidad de las variantes consideradas (Tabla S5 , Anexo)

- **Variante 3**, en el 50% (22 péptidos), con residuos hidrofóbicos en posición +2 o +3
- **Variante 4**, en el 2% (1 péptido) con un residuo [YFH] en alguna de las posiciones variables del *core* y ninguno que únicamente presente [DE] en la posición -1del *core*
- **Variante 5**, en el 11% (5 péptidos) de los *hits* se detectan los residuos del *core* del SLiM.

De los 55 péptidos *hit* de p107 en los que se detectó el SLiM LxCxE se identificaron (Figura 3.3B):

- **Variantes 1 y 2**, en el 36% (20 péptidos) de los *hits*, con características que le confieren la mayor afinidad de las variantes consideradas (Tabla S6 , Anexo)
- **Variante 3**, en el 51% (28 péptidos), con residuos hidrofóbicos en posición +2 o +3
- **Variante 4**, en el 2% (1 péptido) con un residuo [YFH] en alguna de las posiciones variables del *core* y ninguno que únicamente presente [DE] en la posición -1del *core*
- **Variante 5**, en el 11% (6 péptidos) de los *hits* se detectan los residuos del *core* del SLiM.

Aquellos péptidos que presentan la expresión regular de las **Variantes 1 y 2**, tienen características que les confieren una mayor afinidad y podrán ser priorizados en futuros ensayos experimentales (Figura 3.3 intersección entre diagramas de Venn).

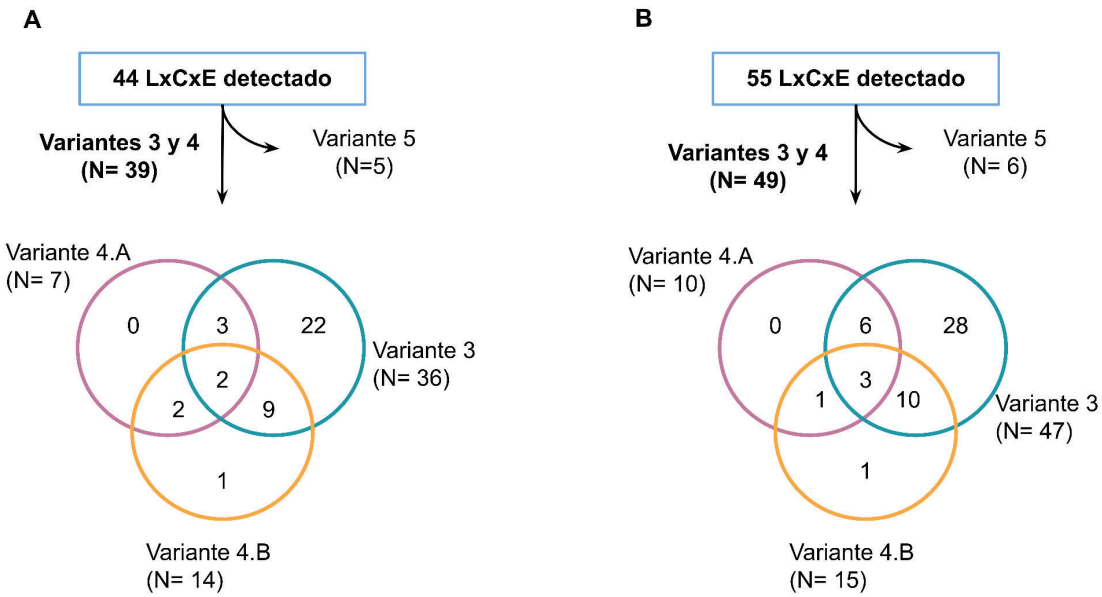


Figura 3.3. Variantes de expresiones regulares del SLiM LxCxE detectadas en péptidos *hit* de las proteínas *pocket*. Del total de péptidos hit con SLiM detectado, se muestran las intersecciones de las Variantes 3 y 4, mientras que los péptidos que presentan la Variante 5 no son considerados para la intersección. La Variante 3 corresponde a péptidos con [WFILYVM] en las posiciones +2 o +3 del core del SLiM (celeste), La variante 4.A corresponde a péptidos con [DE] en la posición -1 del core del SLiM (rosa) y la Variante 4.B a péptidos con [FYH] en alguna de las posiciones variables del ***core*** del SLiM (amarillo). La Variante 1 corresponde a la intersección de las variantes 3, 4.A y 4.B. Las Variantes 2 corresponden a la intersección de la variante 3 con 4.B (2.A), 3 con 4.A (2.B) y 4.A con 4.B (2.C). **A:** Péptidos *hit* de Rb con SLiM LxCxE. **B:** Péptidos *hit* de p107 con SLiM LxCxE.

Finalmente, para Rb y p107 se identificaron péptidos respectivamente que poseen las características de mayor afinidad del SLiM LxCxE (Tabla 3.2), correspondientes a la **Variante 1**. Para Rb, se identificaron dos péptidos correspondientes a las proteínas CPSF7 y ASB3 cuyas secuencias presentan un D en la posición -1, Y en la primera posición variable y un residuo F o W en la posición +2 o +3 del SLiM. Para CPSF7 la presencia de I en vez de L en la posición 1 podría disminuir la afinidad de unión.

Para p107, tres de los péptidos *hit* pertenecientes a la **Variante 1** presentan un residuo E en la posición -1, F o H en la primera posición variable e I en la posición +2. Los péptidos pertenecen a las proteínas UBP10 y KIF15. Esta última posee dos péptidos que son *hits* solapados, que abarcan el mismo segmento de expresión regular (Tabla 3.3). La presencia de T en la posición central (LxTxE) en UBP10 podría disminuir la afinidad de unión.

Tabla 3.3. Péptidos *hit* de las proteínas *pocket* cuya secuencia se identifica con la Variante 1 del SLiM LxCxE.

Proteína	Proteína Pocket	Secuencia*	Localización Celular
CPSF7	Rb	GVDLI**D**I**Y**A**D**E**E**F**F**NQD	Núcleo
ASB3	Rb	GADP**D**L**Y**C**N**E**D**S**W**QLP	Núcleo y citoplasma
UBP10	p107	GTATNGV**E**L**H**T**T**E**S**I**I**D	Núcleo y citoplasma
KIF15	p107	STQMQ**E**L**F**S**S**E**R**I**I**DWT Q**E**L**F**S**S**E**R**I**I**DWTKQQE	Citoplasma

*Secuencias con core del SLiM LxCxE (negro), residuo ácido en posición -1 (rojo), residuo hidrofóbico en posición +2 o +3 (azul).

Es importante resaltar que tanto para Rb como para p107, se observó un alto número de péptidos *hit* con residuos hidrofóbicos en las posiciones +2 o +3 del centro del SLiM (**Variante 3**): 82% (36 de 44 péptidos) y 86% (47 de 55 péptidos), respectivamente. Esto resalta la relevancia funcional de este residuo en la interacción con las proteínas *pocket* (ver Figura 1.6, Sección 1.4.2).

3.2.2. Análisis de la variabilidad de secuencia del SLiM E2F presente en los péptidos *hits*.

Para el SLiM E2F se continúa utilizando la expresión regular de secciones anteriores que fue definida en base a la expresión regular de ELM [36] (ver Sección 1.4.2) junto con el conocimiento existente sobre los determinantes y posiciones que modulan la afinidad de este SLiM:

[IVLMA].[NQ**DE**][IVLMAFYW][IVLMA**FY**W]

De las cinco posiciones, la primera y segunda posición se mantienen igual que lo definido en ELM [36] (ver Sección 1.4.2), con los residuos [IVLMA] seguido de una posición variable (Tabla 3.4) (ver Tabla 7.3, Sección 7.4.2). En la segunda posición fija del SLiM E2F, se consideran N y Q además de los definidos D y E por su capacidad de establecer puentes de hidrógeno con residuos del dominio *pocket* como en E2F2 [33,38]. La cuarta y quinta posiciones fijas del SLiM, se consideraron

los hidrofóbicos [IVAW] además de los definidos por ELM [36], en ambas posiciones con el fin de ampliar la búsqueda a péptidos que presenten residuos similares.

De acuerdo a la evidencia estructural (PDB 1N4M y 2R7G) [33,38], se definieron variantes del SLiM E2F que representan características de alta afinidad de interacción del SLiM con los dominios *pocket* (Tabla 3.4). La **Variante 4,** de expresión regular de los residuos centrales, es lo más inclusiva posible, mientras que las **Variantes 2** y **3**, con residuos de preferencia en la quinta y tercera posición respectivamente, representan variantes que suponemos confieren mayor afinidad de unión debido a la interacción y estabilización de la hélice anfipática en el bolsillo E2F. La **Variante 1** es la combinación de las Variantes 2 y 3 y puede aumentar la afinidad en mayor medida que las otras tres variantes consideradas. Todos estos residuos están presentes en SLiMs que son verdaderos positivos (E1A de distintas clases de Adenovirus, y E2F1-5) y son estéricamente compatibles con el bolsillo de unión, favoreciendo la interacción [33,38]. Si bien residuos flanqueantes hacia los extremos N- y C-terminal del SLiM colaboran con el correcto posicionamiento del SLiM y favorecen la unión, la información disponible para este SLiM no es tan amplia como para el SLiM LxCxE y no se consideran para este análisis.

Tabla 3.4. Definición de variantes del SLiM E2F.

Variante	Características[#]	Expresión regular
1	Res. ácido en tercera posición, res. F/Y en quinta posición	[IVLMA]**.[DE]**[IVLMAFYW]**[FY]**
2	Res. F/Y en quinta posición	[IVLMA].[NQDE][IVLMAFYW]**[FY]**
3	Res. ácido en tercera posición	[IVLMA]**.[DE]**[IVLMAFYW][IVLMAFYW]
4	Res. centrales	[IVLMA].[NQDE][IVLMAFYW][IVLMAFYW]

[#]Listado de las cuatro categorías en orden decreciente de afinidad propuesto.

De un total de 103 péptidos *hit* de Rb en los que se detectó el SLiM E2F, se identificaron las características correspondientes a (Figura 3.4 A):

- **Variante 1**, en el 44% de los *hits* (45 péptidos), con características que le confieren la mayor afinidad de las variantes consideradas (Tabla S7 , Anexo),
- **Variante 2**, en el 6% de los *hits* (6 péptidos),
- **Variante 3**, en el 34% de los *hits* (35 péptidos),
- **Variante 4**, en el 17% de los *hits* (17 péptidos).

En el caso de los 18 péptidos hit de p107 con el SLiM E2F se identificaron las características de (Figura 3.4 B):

- **Variante 1**, en el 28% de los *hits* (5 péptidos) (Tabla S8 , Anexo),
- **Variante 2**, en el 6% de los *hits* (1 péptido) (Tabla S8 , Anexo),
- **Variante 3**, en el 50% de los *hits* (9 péptidos),

- **Variante 4**, en el 17% de los *hits* (3 péptidos).

En ambos casos, los péptidos en los que se detectó la *regex* de la **Variante 1** serán considerados de mayor confianza en la priorización de candidatos para ensayos experimentales (en rojo Figura 3.4, arriba).

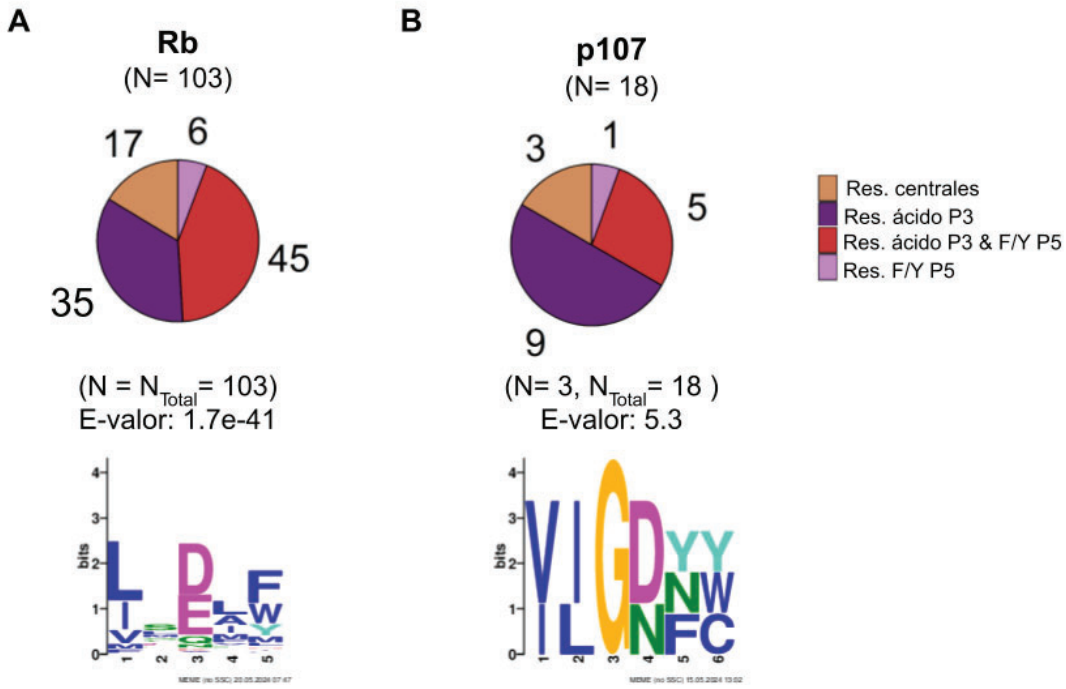


Figura 3.4. Análisis de expresiones regulares de péptidos *hit* con SLiM E2F detectado. A: Distribución de péptidos *hit* de Rb conteniendo variantes del SLiM E2F (*arriba*) y logo de secuencia de los péptidos con SLiM E2F detectado (*abajo*). **B:** Distribución de péptidos *hit* de p107 conteniendo variantes definidas para la expresión regular del SLiM E2F (*arriba*) y logo de secuencia de los péptidos con SLiM E2F detectado (*abajo*). En ambos casos, por fuera del gráfico de torta, se indica el número de péptidos en los que se detectaron las Variantes del SLiM E2F y el número total de secuencias con el SLiM E2F para cada proteína *pocket* (N) (*arriba*). La Variante 2 corresponde a péptidos con residuos [FY] en la quinta posición del core del SLiM (rosa), La variante 3 a péptidos con residuos [DE] en la segunda posición fija del core del SLiM E2F (violeta), la Variante 1 a la combinación de estas dos (rojo) y la Variante 4 a la expresión regular del core más amplia (naranja). Los logos de secuencia fueron construidos en el sitio web MEME [52] utilizando las secuencias de péptidos *hit* de las proteínas *pocket* con SLiM E2F detectado (abajo). Por encima del logo se detalla el número de secuencias alineadas por MEME (N), el número total de secuencias que se ingresaron en MEME (N_{Total}) y el E-valor del alineamiento. El tamaño de cada letra es directamente proporcional a la frecuencia de aparición del residuo en la posición indicada en el eje x. Los colores de cada letra son de acuerdo a características fisicoquímicas de los aminoácidos.

Se construyó un logo de secuencia utilizando las 103 secuencias de los péptidos *hit* de Rb y 18 secuencias de p107 en los que se detectó el SLiM E2F con el fin de identificar SLiMs sobrerrepresentados utilizando la herramienta MEME [52]. El logo de secuencia construido a partir de los datos de Rb muestra la representación de la **Variante 1** del SLiM E2F, con un E-valor de 1.7e-41, que resulta significativo dado que se encuentra por debajo del E-valor de 0.005 (Figura 3.4A, abajo).

Notablemente, en el mismo análisis realizado con los datos de p107 no se identifica un patrón sobrerrepresentado en el logo de secuencias de la proteína *pocket*, donde MEME alineó sólo tres de

las 18 secuencias y el E-valor fue de 5,3 indicando que el patrón obtenido es poco significativo (Figura 3.4 B, abajo).

En resumen, los péptidos *hits* con SLiMs E2F están enriquecidos en las Variantes 1, 2 o 3 con un residuo ácido [DE] y/o aromático [FY] en la segunda y quinta posición fija del SLiM, lo que se corresponde con variantes de alta afinidad del SLiM E2F [38][33],[38]. Estas observaciones sugieren que muchos de los SLiMs presentes en los *hits* de ProP-PD corresponden a variantes de alta afinidad.

3.3. Patrones de secuencia en los péptidos *hit* sin SLiMs conocidos

Por último, con el objetivo de detectar SLiMs novedosos, se analizó por separado los péptidos *hits* en los cuales no se detectó una expresión regular conocida para ninguno de los SLiMs E2F ni LxCxE (Figura 3.5).

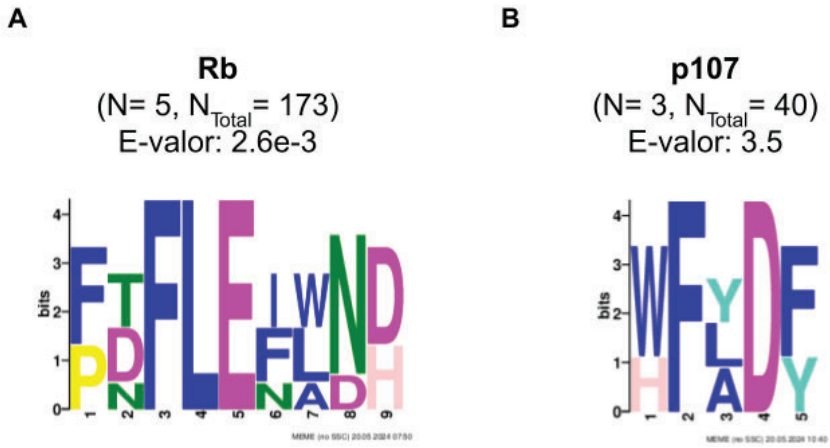


Figura 3.5. Logos de secuencia de péptidos *hit* de la proteínas *pocket* sin SLiM detectado. Representación de residuos por posición presentes en secuencias de péptidos que fueron *hit* de las proteínas *pocket* en el ensayo ProP-PD en los que no se detectaron expresiones regulares de los SLiMs LxCxE y E2F, utilizando la herramienta MEME [52]. El tamaño de cada letra representa la frecuencia de aparición del residuo en la posición indicada en el eje de abscisas. Los colores de cada símbolo corresponden al esquema de colores que MEME utiliza de acuerdo a similitudes bioquímicas de los aminoácidos. **A:** Logo de secuencias de cinco de las 173 péptidos *hit* de Rb sin SLiM detectado. **B:** Logo de secuencias de tres de los 40 péptidos *hit* de p107 sin SLiM detectado.

Para la proteína Rb, MEME detectó en los péptidos *hit* sin SLiM identificado por expresiones regulares, un patrón correspondiente al alineamiento de cinco péptidos provenientes de tres proteínas distintas (Tabla 3.4). Cuatro de estos péptidos, aunque están solapados, poseen una expresión regular similar a la del SLiM E2F que va desde la posición tres a la siete en el logo (Figura 3.5 A). Si bien la F de la primera posición del SLiM, no cumple con la expresión regular canónica de E2F definida en la base de datos ELM, sí lo hacen las posiciones restantes. Este SLiM podría corresponder a una variante novedosa del SLiM E2F, llamada aquí *E2F-like*.

Tabla 3.4. Detalle de péptidos *hit* de Rb sin SLiM detectado alineadas según MEME.

ID Uniprot	Secuencia[#]	SLiM
PRUN2_HUMAN	**PT<u>FLEIW</u>ND**SVDGDSF DRKT**PT<u>FLEIW</u>ND**SVD	E2F-*like*
ZN865_HUMAN	SYP**FD<u>FLEFL</u>NH**QRFE VHFQSYP**FD<u>FLEFL</u>NH**	E2F-*like*
STRN3_HUMAN	VLET**FN<u>FLENA</u>DD**SDE	E2F *reverso*

[#]Se resalta en la secuencia el patrón identificado por MEME y se subrayan los SLiMs similares al SLiM E2F (E2F-like) [52].

Por el contrario, el patrón detectado por MEME en los péptidos *hit* de p107 sin SLiM detectado no fue significativo (E-valor = 3.5) y no observó representación de variantes conocidas de los SLiMs LxCxE o E2F (Figura 3.5 B).

El análisis de SLiMs sobrerrepresentados entre los péptidos que fueron *hit* de Rb sin SLiM reveló una variante novedosa del SLiM E2F (*E2F-like*). Las instancias conocidas del SLiM E2F presentan en la primera posición residuos [LIV], provenientes de las proteínas E2F [38] y E1A de adenovirus [33,38]. El SLiM *E2F-like* posee una fenilalanina en la primera posición. Esta variante es detectada en dos proteínas diferentes (PRUN2_HUMAN y ZN865_HUMAN), una de las cuales fue previamente reportada como interactor en ensayos de proteómica (PRUN2_HUMAN) [48] (ver Tabla 2.3, Sección 2.3). En el caso de la proteína STRN3_HUMAN (Tabla 3.4), MEME identifica el patrón '**F.E**'. Mientras que para el SLiM E2F la expresión regular es ϕ.[DE]ϕϕ, donde ϕ es una posición hidrofóbica, en este caso el patrón observado es ϕϕ[DE].ϕ (Tabla 3.4). Este podría ser un SLiM E2F "invertido" (E2F-reverso) representando un nuevo modo de interacción de la hélice anfipática.

El análisis de *hits* en los que se detectó el SLiM E2F indica la presencia de variantes de alta afinidad de unión al dominio *pocket*. Además, identificamos *hits* conteniendo una F en la primera posición fija (SLiM *E2F-like*) y un E2F invertido (SLiM E2F reverso) que podría representar un nuevo modo de unión. Esto sugiere que ProP-PD puede ser una herramienta poderosa para la detección de variantes de SLiMs conocidos y de SLiMs novedosos de unión a las proteínas *pocket*.

3.4. Conclusiones de patrones de secuencia observados en los péptidos hits

La detección de expresiones regulares del SLiM LxCxE presentes en péptidos *hit* de Rb y p107 reveló que más del 80% presentan residuos hidrofóbicos en la posición +2 o +3 con respecto a los residuos centrales del SLiM. Este resultado es consistente con reportes que proponen la posición hidrofóbica +2/+3 como una posición fija del SLiM [34]. Por otro lado, se observó un enriquecimiento del residuo S en la posición central del SLiM para *hits* de p107, lo que se

corresponde con ensayos experimentales que reportan que el dominio p107 tiene preferencia por el SLiM LxSxE y con la ausencia de interactores de Rb con esta variante. [22,34,50].

Como se discutió previamente, la mayor representación del SLiM E2F en el total de péptidos *hit* de Rb, probablemente, sea debida a la baja recuperación de péptidos con SLiM LxCxE, como consecuencia de mutaciones C→A en la posición central del SLiM, que disminuye la afinidad de unión [40]. En contraste, se recuperaron muy pocos péptidos *hit* con SLiM E2F utilizando el dominio de p107 como carnada, de los cuales en el 56% también se detectó al SLiM LxCxE y no fue posible identificar un SLiM E2F representado empleando la herramienta MEME [52]. Esta menor recuperación de *hits* con SLiM E2F en p107 probablemente se deba a la mayor preferencia y enriquecimiento en SLiMs LxSxE, dado que ProP-PD es un ensayo donde la mayor preferencia por un SLiM puede competir la unión de un segundo SLiM.

Se identificó además, un SLiM similar al SLiM E2F (*E2F-like*) en dos proteínas, con un residuo F en la primera posición fija. Una de estas proteínas fue reportada como interactora en ensayos de proteómica. [48]. Este SLiM podría ser una variante novedosa del SLiM E2F para la cual no existe evidencia experimental de interacción con las proteínas *pocket* reportada hasta la fecha. Por último, se identificó un SLiM E2F invertido en la proteína STRN3_HUMAN que podría representar una forma de interacción novedosa no reportada hasta la fecha.

Teniendo en cuenta la importancia de los determinantes de unión de los SLiMs LxCxE y E2F analizados en este capítulo, las variantes de las expresiones regulares presentadas aquí serán utilizadas para priorizar péptidos *hit*s para su validación experimental.

Capítulo 4: Análisis de parámetros estructurales para filtrado y priorización de péptidos *hit*

En este capítulo, con el fin de establecer una estrategia de filtrado y priorización de péptidos *hit*, se analizaron parámetros estructurales que permitan caracterizar los péptidos utilizando tres herramientas:

1. Accesibilidad relativa al solvente medida en modelos predichos por AlphaFold2 [54] permitiendo puntuar el grado de exposición del péptido para interactuar con el dominio globular;
2. Predicción del grado de desorden utilizando IUPred [55], que permitirá estimar el grado de desorden que presentan los péptidos, priorizar los péptidos con mayor valor de IUPred y filtrar aquellos que posean valores bajos;
3. Detección de dominios Pfam [56], permitiendo determinar el número de residuos de los péptidos *hits* que se encuentran dentro de un dominio Pfam, que en su mayoría son globulares.

Para priorizar péptidos *hit* a ser ensayados experimentalmente, es deseable que cumplan con tres criterios. Primero, un alto grado de exposición al solvente, lo que se refleja en un valor alto de RSA. Segundo, un alto grado de desorden, es decir, un valor alto de IUPred. Finalmente, no deben solaparse con dominios Pfam, ya que probablemente estén formando parte de un dominio globular y por lo tanto, no estén disponibles para interactuar con un dominio de unión.

Si bien para la construcción de la biblioteca HD2 [40] se evaluó previamente algunos de estos parámetros para aumentar la probabilidad de que los péptidos correspondan a IDRs, la selección de péptidos se basó, en primer lugar, en la accesibilidad relativa al solvente (RSA, ***R****elative* ***S****olvent* ***A****ccessibility*) [57] obtenido a partir de proteínas con estructuras resueltas o, en su defecto, proteínas homólogas con estructuras resueltas. De no cumplirse estos criterios, se recurrió al algoritmo IUPred [55] para la predicción del desorden estructural, conservando péptidos cuya secuencia es predicha como desordenada. Sin embargo, al momento del diseño de la biblioteca HD2, menos del 17% de los residuos del proteoma humano tenían una estructura resuelta asociada [58] y la herramienta de predicción de estructuras, AlphaFold2, no estaba disponible.

En el año 2020, AlphaFold2 superó ampliamente a los predictores de estructura en la 14va edición de la competencia mundial para la predicción de estructura proteica, CASP (***C****ritical* ***A****ssessment of techniques for protein* ***S****tructure* ***P****rediction*) [59]. Hasta el año 2021, AlphaFold2 logró una cobertura del 58% de los residuos del proteoma humano y un 38% predichos con alta confianza [58]. Actualmente, AlphaFold2 es confiable para evaluar el RSA de péptidos sin estructura [54,60].

Por otro lado, si bien se utilizó IUPred [55] para complementar la falta de estructuras disponibles durante la construcción de la biblioteca, IUPred presenta limitaciones al predecir segmentos desordenados en un contexto muy ordenado (ver Sección 7.5.2). Por lo tanto, incorporar la determinación de valores de RSA permite una confianza mucho mayor en la priorización y selección de los péptidos *hits*.

4.1. Accesibilidad Relativa al Solvente

Para evaluar si los péptidos *hit* se encuentran accesibles y expuestos para interactuar con las proteínas *pocket*, se obtuvo un valor de accesibilidad relativa al solvente (RSA) a partir de estructuras modeladas en AlphaFold2 [54] para cada proteína a la cual pertenecen los péptidos *hits*. El parámetro RSA toma valores entre cero y uno, y se consideraron péptidos con valores de RSA mayor a 0,4 como accesibles y expuestos para interactuar con las proteínas *pocket* (Figura 4.1).

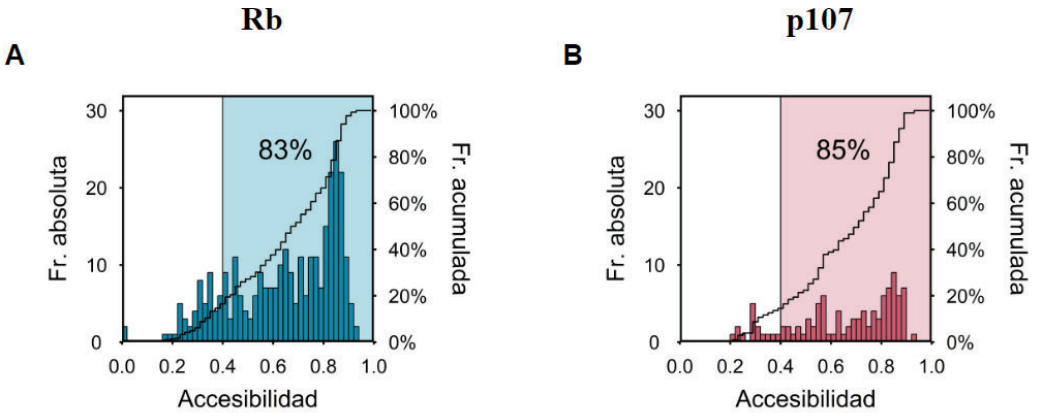


Figura 4.1. Accesibilidad de péptidos que fueron *hit* de las proteínas *pocket* en el ensayo ProP-PD. Frecuencia absoluta de péptidos que fueron *hit* de los dominios *pocket* de Rb **(A)** y p107 **(B)**. En el eje *y* de la derecha, figura para ambos casos la frecuencia acumulada de péptidos, indicando que a medida que el valor de accesibilidad aumenta, también lo hace el número de péptidos que se encuentran expuestos para interactuar.

En el caso de los péptidos de interacción con el dominio *pocket* de Rb, el 83% (257 de 308) de los mismos tienen un valor de RSA mayor o igual a 0,4 (Figura 4.1A); mientras que en el caso de p107 el 85% (88 de 103) igualan o superan este umbral, indicando que se encuentran accesibles para interactuar con blancos proteicos (Figura 4.1B).

Criterio de filtrado y priorización utilizado: El parámetro RSA se utilizará como filtro de corte. Los péptidos que posean un valor de RSA ≥ 0.4 serán incluidos. Se utiliza un valor más permisivo que el de 0.581 reportado en la literatura para la predicción de sitios de unión en péptidos de 25 residuos [60] con el objetivo de no descartar péptidos con valores de RSA aceptables, que puedan representar

nuevos interactores. Los péptidos que posean un valor de RSA ≥ 0.4, serán priorizados en primer lugar en base a este valor.

4.2. Predicción del desorden utilizando el algoritmo IUPred

El algoritmo IUPred predice, a partir de una secuencia, IDRs en base a la energía de un residuo en el contexto de aminoácidos en el que se encuentra. El valor de IUPred toma valores entre 0 y 1. Un valor superior a 0.5 indica una alta propensión al desorden y un valor menor a 0.5 indica una alta propensión al orden. Si bien el estándar es considerar valores por encima de 0,5 como residuos desordenados, pueden utilizarse umbrales menores, más inclusivos. Dado que IUPred no es la única herramienta utilizada en este análisis y su algoritmo presenta algunas desventajas en el análisis de las secuencias (ver Sección 7.5.2), se definió un umbral de 0,4 para priorizar péptidos *hit* de manera más abarcativa y seleccionar entre ellos a los mejores candidatos. La predicción de IUPred se realizó sobre las secuencias enteras de las proteínas a las cuales pertenecen los péptidos *hits,* ya que esta herramienta considera el contexto de la secuencia para calcular la puntuación de cada residuo. Luego, para cada péptido *hit* se calculó el porcentaje de residuos desordenados (IUPred ≥ 0,4) sobre el total de los 16 residuos de su secuencia (Figura 4.2).

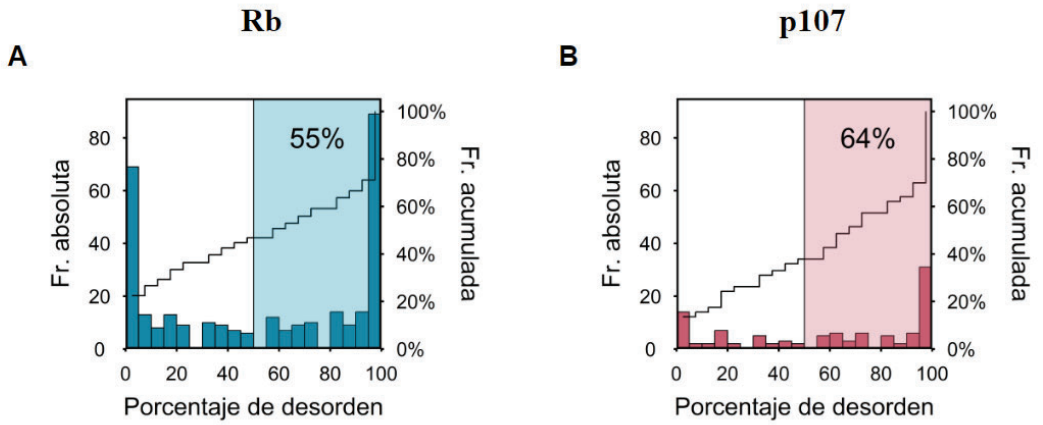


Figura 4.2. Predicción del grado de desorden de péptidos que fueron *hit* de las proteínas *pocket* utilizando IUPred. Frecuencia absoluta de péptidos que fueron *hit* de los dominios *pocket* para Rb **(A)** y para p107 **(B)**. En el eje *y* de la derecha, se indica para ambos casos la frecuencia acumulada de péptidos, indicando que a medida que el porcentaje de desorden aumenta, también lo hace el número de péptidos con un alto grado de desorden.

Para ambas proteínas *pocket*, se observan péptidos que poseen un bajo porcentaje de residuos desordenados (Figura 4.2). Estos péptidos podrían corresponder a extremos de dominios globulares, o por ejemplo, a *loops* que se encuentran entre regiones transmembrana y debido al contexto ordenado, IUPred los puntúa con un valor bajo.

Al utilizar el dominio *pocket* de Rb como carnada, el 55% de los péptidos *hit* (170 de 308)

poseen un alto grado de desorden (≥ 50% de residuos desordenados) (Figura 4.2A). Para el caso de p107, el 64% de sus péptidos *hits* (66 de 103) se consideran altamente desordenados (Figura 4.2B). Se espera que estos péptidos se encuentren expuestos al solvente y disponibles para interactuar con un dominio globular.

Criterio de priorización utilizado: Dado que IUPred ya fue utilizado en la construcción de la biblioteca HD2, no se utilizará IUPred como filtro de corte, sino que se priorizarán aquellos péptidos *hit* que presenten valores promedio de IUPred mayores, ya que se espera que un contexto altamente desordenado sea más favorable para la interacción con un dominio globular. Los péptidos *hit* con un valor de IUPred menor o igual a 0,4 serán señalados con advertencias al momento de priorizarlos. Este valor es mucho más permisivo que el valor de 0.5 utilizado en la literatura.

4.3. Detección de dominios Pfam

La base de datos Pfam reúne información de familias de proteínas representadas por alineamientos múltiples de secuencia y modelos ocultos de Markov (HMM) que en su gran mayoría, aunque no siempre (ver Sección 7.5.3), se corresponden con dominios globulares [56]. Si bien el 80% de los SLiMs se encuentran en IDRs, también pueden estar en loops flexibles de dominios globulares [2].

Se realizó la detección de dominios Pfam en las secuencias enteras de las proteínas a las que pertenecen los péptidos *hits*. Luego, se cuantificó el número de residuos que los péptidos *hit* comparten con un dominio Pfam (Figura 4.3).

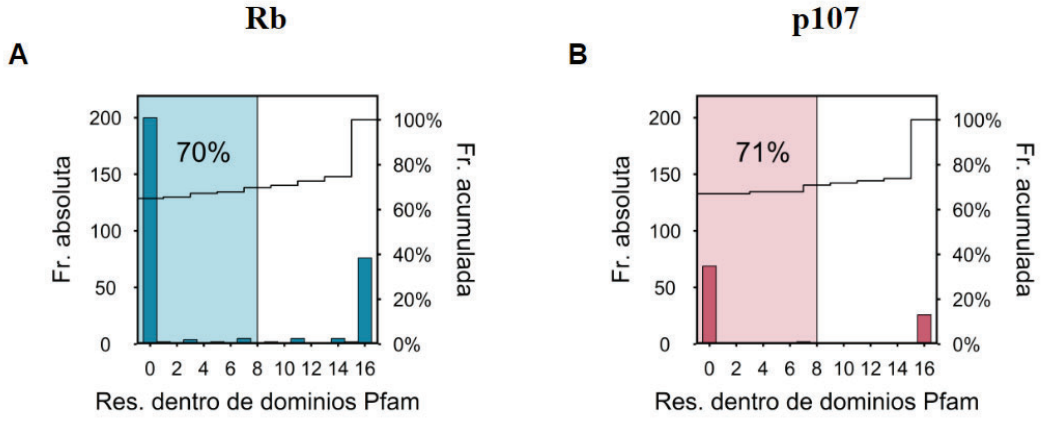


Figura 4.3. Detección de dominios Pfam en péptidos que fueron *hit* de las proteínas *pocket* en el ensayo ProP-PD. Frecuencia absoluta de péptidos que fueron *hit* de los dominios *pocket* para Rb (A) y para p107 (B). En el eje *y* de la derecha, figura para ambos casos la frecuencia acumulada de péptidos, que muestra que la mayoría de los péptidos no comparten residuos con dominios Pfam y a medida que éste número aumenta, la curva se mantiene estable.

Para este análisis se considera que un péptido *hit* está solapado con un dominio Pfam si más de la mitad de su secuencia, es decir, más de ocho residuos, coinciden con dicho dominio. En base a esto, el 70% (215 de 308) de los péptidos *hit* de Rb como carnada, no se encuentra solapado con dominios Pfam (Figura 4.3A), mientras que el 71% (73 de 103) en el caso de p107, cumplen con esta condición (Figura 4.3B).

Criterio de priorización de candidatos: La identificación de péptidos que presentan solapamiento con dominios de la base de datos Pfam, podría sugerir que estos pertenecen a dominios globulares. Sin embargo, Pfam también incluye dominios desordenados que han sido conservados a lo largo de la evolución y es posible encontrar SLiMs en *loops* flexibles que conectan dominios globulares [2]. Por lo tanto, esta herramienta no será utilizada como filtro de corte sino que se priorizarán los péptidos *hit* que compartan ocho o menos residuos de su secuencia con algún dominio Pfam, mientras que aquellos que compartan nueve o más, serán identificados con advertencias.

4.4. Conclusión del análisis de parámetros estructurales para filtrado y priorización de péptidos *hit*.

Evaluar los parámetros estructurales nos permite interpretar con mayor confianza si se espera que los péptidos estén accesibles para interactuar con los dominios *pocket*. Las herramientas empleadas en esta sección deben considerarse en conjunto, ya que ninguna es concluyente evaluada de manera individual. Sin embargo, los resultados indican que la mayoría de los péptidos *hit* presentan un alto grado de desorden en su secuencia, son expuestos y accesibles para interactuar con blancos proteicos, y no se solapan con dominios Pfam. Esto era en parte esperado, ya que estos criterios fueron los aplicados al diseño de la biblioteca HD2 de la cual provienen, aunque herramientas como la predicción de estructuras con AlphaFold2 [2,54] no se encontraba disponible al momento de la construcción de la biblioteca y permitió una estimación más confiable del valor de RSA. Esta herramienta fue crucial para diseñar una estrategia de filtrado y priorización que permita al laboratorio optimizar el uso de tiempo y recursos en ensayos experimentales.

Por lo tanto, se utilizará primero la Accesibilidad Relativa al Solvente (RSA) para filtrar péptidos que se encuentren por debajo de un valor de 0,4. Luego, se evaluará el valor de IUPred y la superposición con dominios Pfam en aquellos péptidos que superen el valor de RSA. Si poseen un valor de IUPred menor a 0,2 (altamente ordenado) y están solapados con un dominio Pfam en más del 50% de su secuencia, también serán filtrados.

Aquellos péptidos que posean un valor de IUPred entre 0,2 y 0.4, y/o más de ocho residuos solapados con un dominio Pfam, serán identificados con advertencias o *warnings* por poder pertenecer a regiones ordenadas y/o a dominios globulares. La lista de péptidos filtrados, será priorizada por los

valores de RSA e IUPred, siendo prioritarios los *hits* con valores más cercanos a uno en ambos parámetros.

Capítulo 5: Estabilidad energética de péptidos *hit*.

En este capítulo, se presenta un análisis de estabilidad de los péptidos que fueron *hit* en ProP-PD, con el fin de identificar dentro del set de datos, candidatos de mayor afinidad predicha para ser priorizados en ensayos experimentales. Para cumplir con este objetivo, se utilizaron matrices FoldX disponibles en el laboratorio de trabajo.

FoldX es un campo de fuerza que estima a partir de una estructura tridimensional o modelo estructural de un complejo, por ejemplo péptido-Dominio Globular, la afinidad de un péptido en términos de la variación de energía libre de Gibbs ($\Delta\Delta G$) que ocurre al sustituir un residuo por otro en el péptido del complejo. El resultado, es una matriz de valores de penalización de sustitución con tantas filas como residuos tiene el péptido y 20 columnas por los 20 aminoácidos [61]. Luego, se normaliza esa matriz a los residuos que se encuentran en el péptido del complejo de manera tal que el péptido del complejo tiene un $\Delta\Delta G = 0$. Una mayor afinidad de un péptido implica que su estabilidad energética es mayor y obtiene una menor penalización de FoldX. A menor valor de FoldX, mayor estabilidad del péptido en el complejo y, por lo tanto, mejor es su interacción; mientras que a mayor valor de FoldX, el péptido no es estable energéticamente en el complejo y no resulta un buen candidato de interacción.

- *Matrices FoldX utilizadas*: Se evaluó la estabilidad energética de péptidos *hit* utilizando matrices FoldX disponibles en el laboratorio de trabajo a partir de complejos de las proteínas *pocket* identificados en Protein Data Bank (PDB). Para el SLiM LxCxE se utilizaron inicialmente las matrices obtenidas de los complejos E7-Rb (PDB: 1GUX) (Tabla S9, Anexo) y el complejo LIN52-p107 (PDB : 4YOS) que contiene la variante LxSxE (Tabla S10, Anexo). Un análisis preliminar del SLiM LxCxE reveló que la matriz 4YOS no refleja adecuadamente la evidencia experimental de mutagénesis sobre las posiciones clave para la interacción del péptido con el dominio *pocket* siendo descartada del análisis (ver Sección 7.6.4) [34,50]. Para el SLiM E2F se utilizaron las matrices provenientes de los complejos E1A-Rb (PDB: 2R7G) (Tabla S11, Anexo)y el complejo E2F2-Rb (PDB: 1N4M) (Tabla S12, Anexo) [62].
- *Sets de datos para benchmarking*: Para evaluar la performance de las matrices FoldX y seleccionar las variantes a utilizar, se utilizaron tres sets de datos, a saber:
    - **Péptidos TP/TN:** Instancias de SLiMs verdaderos positivos (***T**rue **P**ositives,* TP) reportadas en ELM [36] para Rb y p107; y variantes de la proteína viral E7 de HPV testeadas en el laboratorio. Todos ellos, conteniendo variantes de los SLiMs LxCxE y E2F. Los SLiMs verdaderos negativos (***T**rue **N**egative,* TN) son las secuencias de los SLiMs TP con las posiciones fijas del SLiM, mutadas por alaninas (A) (Tablas S13,

S14, S15, S16, Anexo).

- **Péptidos *hit* de ProP-PD testeados experimentalmente:** péptidos que fueron *hit* en el ensayo ProP-PD que el laboratorio validó experimentalmente utilizando técnicas *in vitro* de interacción proteína-proteína utilizando dominios *pocket* de Rb y p107 (Tablas S17, S18, S19 y S20, Anexo). Incluye péptidos que contienen un SLiM.
- **Péptidos *hit* de ProP-PD:** la totalidad de péptidos que fueron *hit* en el ensayo ProP-PD para los dominios *pocket* de Rb y p107. En cada caso, se dividieron en dos grupos: péptidos en los que se detectó un SLiM y péptidos en los que no.

5.1. Evaluación del SLiM LxCxE

5.1.1. Evaluación de matrices FoldX en interactores conocidos (TP) con SLiM LxCxE

Para evaluar si las matrices de FoldX permiten clasificar péptidos como estables o inestables, se utilizaron primero para evaluar un conjunto de datos de interactores conocidos (*true positives*, TP) reportados en la base de datos de motivos lineales, ELM [36]. Utilizando la matriz FoldX de 1GUX, se escanearon 42 péptidos TP de la proteína Rb (Tabla S13, Anexo) y 33 TP de p107 (Tabla S14, Anexo) conteniendo el SLiM de interacción LxCxE.

Se definieron dos variantes de la matriz 1GUX (Tabla 5.1) para profundizar el análisis de secuencias estables, dado que la matriz en la primera posición (aspártico, D) no penaliza diferente a los aminoácidos y permite la sustitución por igual por cualquier aminoácido. La primera variante considerada fue la matriz denominada **1GUX_9** (`DLYCYEQLN`) que incluye los residuos de la posición uno a nueve. La segunda variante, **1GUX_8** (`LYCYEQLN`), incluye a los residuos de la posición dos a nueve de la matriz 1GUX.

Tabla 5.1. Posiciones y residuos de la matriz 1GUX.

	1GUX								
Posición#	1	2	3	4	5	6	7	8	9
Residuo	D	**L**	**Y**	**C**	**Y**	**E**	Q	L	N

#Nueve posiciones de la matriz 1GUX [5], basada en la estructura cristalográfica del complejo formado por el péptido de la proteína viral E7 de HPV con el dominio *pocket* de Rb.
Rojo: C***ore*** del SLiM LxCxE.

Con el fin de analizar la capacidad de las dos variantes de identificar a los TPs del SLiM LxCxE, se escanearon las secuencias de los interactores TP registrando la ocurrencia de la expresión regular del SLiM (`[IL].[CAST].E`), y los mínimos de FoldX, marcando con un círculo amarillo

las subsecuencias positivas para la *regex* y con un mínimo de FoldX (detección positiva) (Figura 5.1). Luego, se registró el porcentaje de detecciones positivas para cada matriz para identificar qué variante de la matriz tiene mejor capacidad para detectar a las subsecuencias TP (Tabla 5.2).

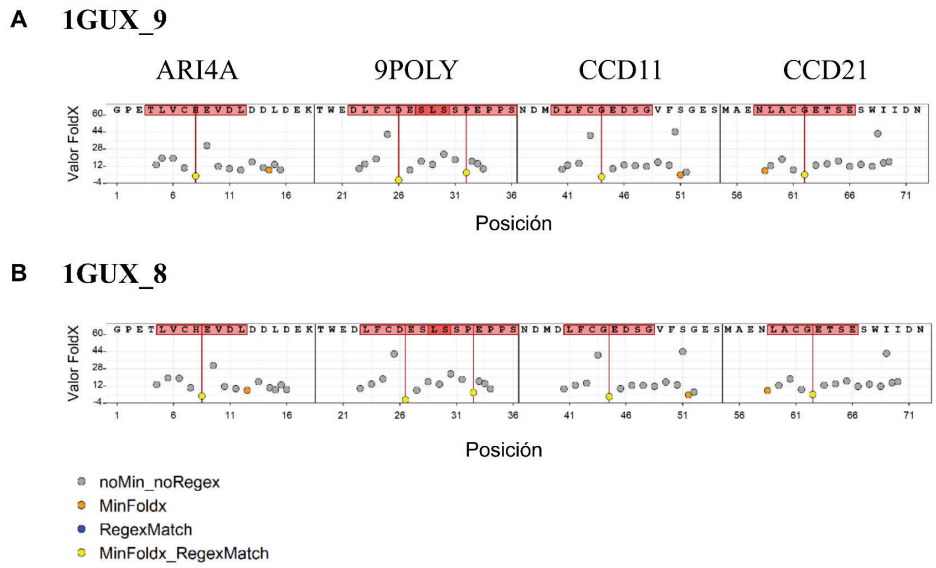


Figura 5.1. Distribución de valores FoldX de péptidos TP con SLiM LxCxE reportados para Rb. Fragmento de gráfico de puntos de péptidos conocidos de interacción con la proteína *pocket* Rb. Los 18 residuos de péptidos que pertenecen a las proteínas ARI4A_HUMAN, 9POLY_HUMAN, CCD11_HUMAN y CCD21_HUMAN fueron escaneados con **A:** 1GUX_9, y **B:** 1GUX_8. El eje *x* indica la posición como el punto medio entre la posición inicial y final de cada sub-secuencia analizada (abajo) y el detalle de residuo de la secuencia peptídica (arriba). Las secuencias se organizaron de manera continua una de otras y se marcó el inicio de una secuencia distinta con una línea vertical negra. Las líneas verticales rojas indican el punto medio de la subsecuencia donde se detectó la expresión regular `[IL].[CAST].E`. Los puntos amarillos indican la detección de una subsecuencia conteniendo al SLiM LxCxE (recuadro rojo), con el valor mínimo de FoldX de todas las subsecuencias escaneadas para ese péptido. El gráfico de puntos de los 42 interactores TP para Rb y 33 para p107 escaneados con 1GUX_9 y 1GUX_8 se encuentra en las Figuras Suplementarias S1 y S2 del apartado Anexo.

Tabla 5.2. Capacidad de detección de TP por variantes de 1GUX

Proteína *pocket*	**TP (N-total)**	**1GUX_9*** (% SLiMs detectados #)	**1GUX_8**** (% SLiMs detectados #)
Rb	42	98	98
p107	33	97	97

% SLiMs detectados: % péptidos con valor mínimo de FoldX en la subsecuencia positiva para la *regex* del SLiM LxCxE.
*1GUX_9: variante de 1GUX de 9 residuos utilizada para el escaneo de secuencias.
**1GUX_8: variante de 1GUX de 8 residuos utilizada para el escaneo de secuencias.

La Figura 5.1 y la Tabla 5.2 muestran que ambas matrices son capaces de identificar dentro de los péptidos a las subsecuencias TP como hits que contienen la expresión regular del SLiM LxCxE y que son las subsecuencias más estables del péptido. Esto se cumple para ambas proteínas *pocket*.

5.1.2. Análisis de *recall* y especificidad de interactores conocidos (TP) con SLiM LxCxE

Con el objetivo de realizar un *benchmarking* que permita definir un valor de corte o umbral de FoldX que distinga secuencias que unen de las que no, se utilizó la lista de TP y se construyeron los verdaderos negativos (*true negative*, TN) mutando las posiciones fijas del SLiM LxCxE a alaninas. Luego, se utilizaron las matrices FoldX para evaluar qué variante de 1GUX distingue mejor estos dos grupos de péptidos. Para ello, se seleccionaron sub-secuencias de los péptidos TP que contengan la expresión regular y en una segunda instancia, que presenten el mínimo valor de FoldX. Se identificó la posición de inicio y fin de la subsecuencia en el péptido TP y se seleccionó la subsecuencia comprendida entre las mismas posiciones de los péptidos TN, de manera que se elige el mismo segmento de los péptidos TP y TN (Figura 5.2 A y Figura Suplementaria S3, Anexo) registrando el valor de FoldX de cada uno (Tabla S13 y S14, Anexo).

Luego, se realizó el cálculo del porcentaje de *recall* (recuperación de TP) y la especificidad (recuperación de TN) de cada variante utilizando distintos valores de FoldX como umbral (Figura 5.2 B y Figura Suplementaria S3, Anexo). Esto permitió evaluar la capacidad de cada matriz de distinguir TP de TN (Tabla 5.3).

El *recall* se calculó de la siguiente manera:

$$Recall = \frac{TP}{TP\ total} * 100$$

donde *"TP"* es el número de péptidos que se encuentran por debajo del umbral y *"TP total"*, la cantidad total de TP. La especificidad se calculó utilizando la siguiente relación:

$$Especificidad = \frac{TN}{TN+FP}$$

donde *"TN"* es el número de péptidos que se encuentran por encima del umbral definido y *"TN+FP"*, el número total de péptidos TN del conjunto de datos comparativos (ver sección 7.6.2).

Tabla 5.3. Métricas aplicadas a variantes de 1GUX en el escaneo de interactores conocidos.

Proteína Pocket	TP (N-total)	TN (N-total)	Métrica[#]	1GUX_9*	1GUX_8**
Rb	42	42	**Recall (%)**	83	90
			Especificidad	1	1
p107	33	33	**Recall (%)**	88	91
			Especificidad	1	1

[#] Valores de *recall* y especificidad utilizando un punto de corte de cinco.
*1GUX_9: variante de 1GUX de 9 residuos utilizada para el escaneo de secuencias.
**1GUX_8: variante de 1GUX de 8 residuos utilizada para el escaneo de secuencias.

El primer resultado observado es que para ambas proteínas *pocket*, el conjunto de péptidos TP poseen valores de FoldX menores que los TN, indicando mayor estabilidad energética de los péptidos TP en comparación a los péptidos TN (Figura 5.2 A y Figura Suplementaria S3, Anexo). Si bien el mayor *recall* se obtiene con un valor FoldX umbral de seis (Figura 5.2 B y Figura Suplementaria S3, Anexo), se utilizará un valor umbral de cinco para restringir la priorización de péptidos a aquellos con un menor valor de FoldX que indican una mayor estabilidad. En esta condición, el uso de la variante 1GUX_8 permitió obtener los valores más altos de *recall* y especificidad (Tabla 5.3).

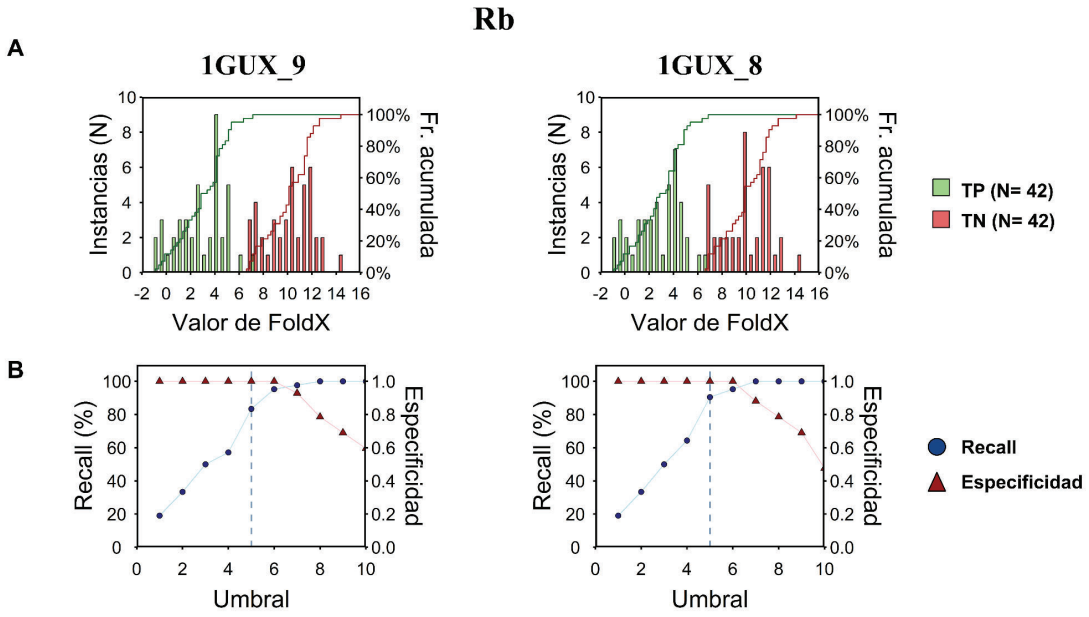


Figura 5.2. Interactores conocidos de Rb escaneados con variantes 1GUX. A: Distribución de valores FoldX de péptidos TP y TN para matrices 1GUX_9 (izquierda) y 1GUX_8 (derecha). Eje Y izquierdo: Número de instancias, Eje Y derecho: porcentaje de frecuencia acumulada de instancias. **B:** R*ecall* (puntos azules) y especificidad (triángulos en rojo) de matrices 1GUX_9 (izquierda) y 1GUX_8 (derecha) para diferentes valores umbral de FoldX. La línea vertical punteada señala el valor de umbral utilizado para establecer comparaciones.

5.1.3. Evaluación de valores FoldX para péptidos *hit* de ProP-PD con SLiM LxCxE testeados experimentalmente

Los siguientes análisis se realizaron con el objetivo de determinar si los valores FoldX son capaces de reflejar la fuerza de unión de *hits* experimentalmente validados del experimento ProP-PD. Se testearon experimentalmente 26 péptidos para Rb (Tabla S17, Anexo) y 22 para p107 (Tabla S18, Anexo) los cuales fueron clasificados según su fuerza de interacción con los dominios *pocket* como péptidos de interacción fuerte (*Strong Positive Binders*, SP), débil (*Weak Positive Binders*, W) y sin interacción (*Non Binders* o *Negative Binders*, N). Los péptidos N son secuencias que poseen la expresión regular del SLiM LxCxE pero sin embargo, no presentaron evidencia de interacción y en este sentido se distinguen de los TN del ensayo anterior, en los cuales la mutación a alanina remueve el SLiM. Por lo tanto, los péptidos N "negativos" de este ensayo presentan una prueba más difícil de

diferenciar de las secuencias que sí presentan evidencia de unión.

Se escanearon las secuencias de péptidos testeados con las matrices 1GUX_9 y 1GUX_8, seleccionando las subsecuencias en las que se detectó la expresión regular y mínimo valor de FoldX (Figura 5.3 A y Figura Suplementaria S4, Anexo). Asimismo se obtuvo el *recall* y la especificidad a diferentes umbrales para definir si existe un umbral que permita clasificar péptidos *hit* de acuerdo a su estabilidad energética (Figura 5.3 A y Figura Suplementaria S4, Anexo). Para esto, se consideraron péptidos positivos a la suma de los SP y W.

La primera observación es que los valores de FoldX caen en un rango similar al de los TPs de ELM (Figuras 5.2 A y Figura Suplementaria S3, Anexo). Sin embargo, los resultados demuestran que, a diferencia de lo observado en TPs, no es posible seleccionar un umbral de FoldX que distinga positivos de negativos, dado que a valores bajos de umbral la especificidad es alta, pero el *recall* no lo es (Tabla 5.4). Los valores FoldX también fueron similares entre péptidos SP y péptidos W.

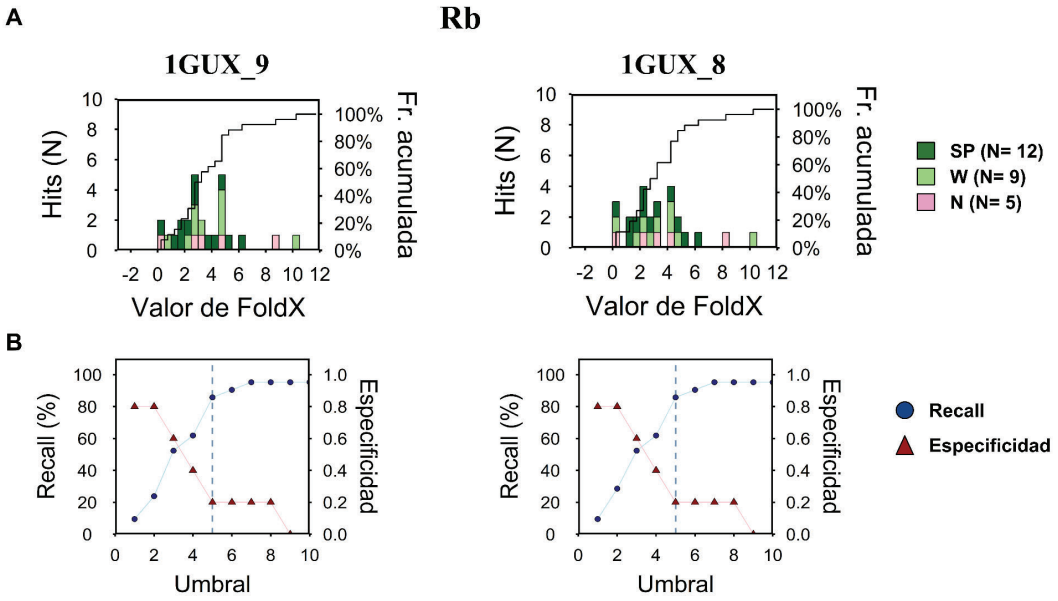


Figura 5.3. Péptidos *hit* en ProP-PD usando Rb como carnada que fueron testeados experimentalmente y escaneados con variantes 1GUX. A: Distribución de valores FoldX de péptidos TP y TN para matrices 1GUX_9 (izquierda) y 1GUX_8 (derecha). Eje Y izquierdo: Número de instancias, Eje Y derecho: porcentaje de frecuencia acumulada de instancias. **B:** R*ecall* (puntos azules) y especificidad (triángulos en rojo) de matrices 1GUX_9 (izquierda) y 1GUX_8 (derecha) para diferentes valores umbral de FoldX. La línea vertical punteada señala el valor de umbral utilizado para establecer comparaciones.

Tabla 5.4. Métricas comparativas aplicadas al escaneo de péptidos *hit* de ProP-PD testeados experimentalmente, con las variantes de 1GUX.

Proteína Pocket	Binders SP+W (N-total)	Non binders (N-total)	Métrica[#]	1GUX_9*	1GUX_8**
Rb	21	5	**Recall (%)**	86	86
			Especificidad	0,2	0,2
p107	16	6	**Recall (%)**	94	94
			Especificidad	0,3	0,3

[#] Valores de *recall* y especificidad utilizando un punto de corte de cinco.
*1GUX_9: variante de 1GUX de 9 residuos utilizada para el escaneo de secuencias.
**1GUX_8: variante de 1GUX de 8 residuos utilizada para el escaneo de secuencias.

En el análisis de escaneo de secuencias para péptidos testeados experimentalmente indica que no es posible determinar un umbral o una variante de matriz 1GUX en el que se observen altos valores de *recall* y especificidad para ninguna de las proteínas *pocket* (Figuras 5.3 B y Figura Suplementaria S4, Anexo). Sin embargo, la alta especificidad a bajos valores de FoldX, sugiere que un valor menor de FoldX (por ejemplo dos) podría utilizarse para detectar péptidos de mayor afinidad.

5.1.4. Evaluación de la *performance* de FoldX sobre todos los péptidos *hit* del ensayo ProP-PD

Con el objetivo de determinar si las matrices FoldX utilizadas son capaces de distinguir si los péptidos que contienen al SLiM LxCxE son más estables que los que no presentan variantes del SLiM, se escanearon secuencias de estos dos grupos de péptidos y se evaluó su estabilidad energética utilizando variantes de 1GUX de ocho y nueve residuos.

De los 308 péptidos *hits* de Rb, 173 no presentaban un SLiM detectado, 32 contenían un SLiM LxCxE, y 12 incluían ambos SLiMs. En cuanto a p107, de los 103 *hits*, 40 no tenían un SLiM detectado, 45 contenían un SLiM LxCxE, y 10 presentaban ambos SLiMs (ver Tabla 3.1, Sección 3.2). En esta sección se analizarán los péptidos con SLiM LxCxE: **44 para Rb y 55 para p107**.

La lista de péptidos *hit* del ensayo ProP-PD para ambas proteínas *pocket* fueron divididos en dos grupos según si se detectó al SLiM LxCxE en su secuencia o no:

1. Péptidos con variantes del SLiM LxCxE que cumplen con la expresión regular `[IL].[CAST].E`
2. Péptidos sin SLiM detectado.

Se escanearon los dos grupos de péptidos con las variantes de 1GUX (Figuras 5.4 y Figura Suplementaria S5, Anexo). Se observa que para ambas proteínas *pocket*, los péptidos con SLiM LxCxE dan distribuciones con valores menores de FoldX cuando se las compara con los péptidos sin este SLiM. Por ende, estas matrices tienen una buena capacidad de distinguir *hits* con SLiM LxCxE.

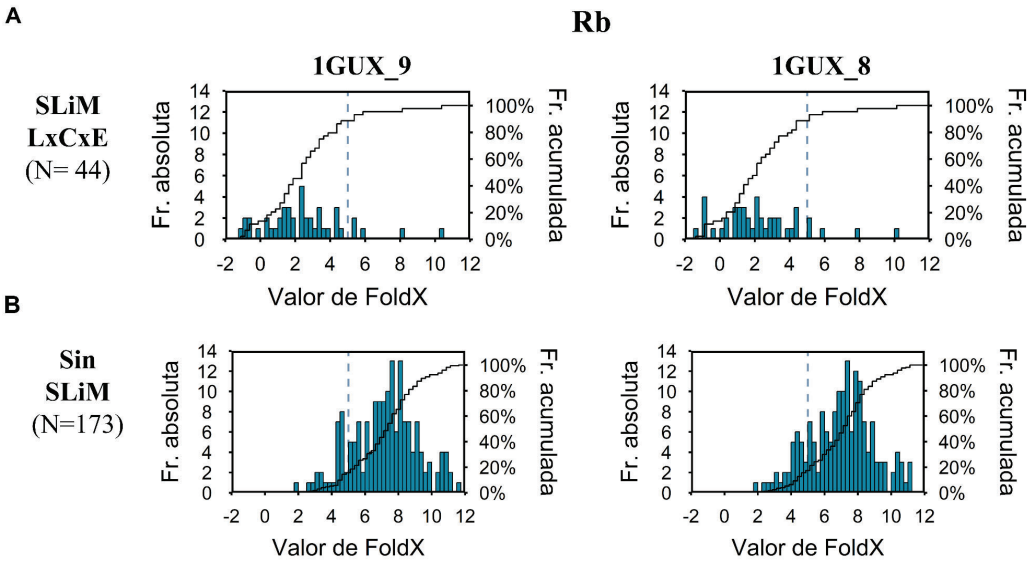


Figura 5.4. Péptidos *hit* de Rb escaneados con variantes 1GUX. A: Distribución de valores FoldX de péptidos para matrices 1GUX_9 (izquierda) y 1GUX_8 (derecha). Eje Y izquierdo: Frecuencia absoluta de instancias, Eje Y derecho: porcentaje de frecuencia acumulada de instancias. La línea vertical punteada señala el valor de umbral utilizado para establecer comparaciones. **A:** Péptidos con SLiM LxCxE detectado **B:** Péptidos sin SLiM LxCxE detectado.

Se calculó como métricas el *recall* y la especificidad utilizando un umbral de cinco para definir si alguna variante de 1GUX es capaz de distinguir ambos grupos (Tabla 5.5). Se consideraron para estas métricas a los péptidos **con SLiM** como **interactores positivos (TP)** y a los péptidos **sin SLiM** como **interactores negativos (TN)**.

Tabla 5.5. Métricas comparativas aplicadas al escaneo de péptidos *hit* de ProP-PD con las variantes de 1GUX.

Proteína Pocket	TP (N-total)	TN (N-total)	Métrica#	1GUX_9*	1GUX_8**
Rb	44	173	Recall (%)	89	89
			Especificidad	0,8	0,8
p107	55	40	Recall (%)	89	96
			Especificidad	0,7	0,6

Valores de *recall* y especificidad utilizando un punto de corte de cinco.
*1GUX_9: variante de 1GUX de 9 residuos utilizada para el escaneo de secuencias.
**1GUX_8: variante de 1GUX de 8 residuos utilizada para el escaneo de secuencias.

Considerando un umbral de cinco de FoldX, se observó el mismo porcentaje de *recall* y misma especificidad para Rb utilizando las dos variantes de 1GUX. Por otro lado, utilizando la variante 1GUX_8 para escanear secuencias *hits* de p107, se obtiene un *recall* más alto aunque la especificidad disminuye en comparación de 1GUX_9 (Tabla 5.5).

Si bien ambas matrices tienen una performance similar, 1GUX_8 mostró un *recall* más alto en comparación con 1GUX_9 en el análisis de interactores conocidos (Tabla 5.3).

Conclusiones generales de la *performance* de las matrices FoldX para el SLiM LxCxE: Se puede concluir que las matrices 1GUX tienen una buena capacidad de distinguir *hits* estables que contienen

SLiMs LxCxE. En base a los resultados presentados, se utilizará 1GUX_8 para evaluar la estabilidad de péptidos *hit* con SLiM LxCxE de Rb y p107 considerando un umbral de FoldX de cinco como punto de corte en la priorización de péptidos candidatos. En base al análisis de péptidos *hit* ProP-PD validados experimentalmente, se concluye que se puede utilizar un umbral de dos para priorizar péptidos de alta afinidad de unión.

5.2. Evaluación del SLiM E2F

5.2.1. Evaluación de matrices FoldX en interactores conocidos (TP) con SLiM E2F.

Se utilizó el conjunto de datos de interactores conocidos, TP, conteniendo variantes del SLiM E2F siguiendo la expresión regular: `[IVLMA].[NQDE][IVLFMYAW][IVLFMYAW]`, para evaluar si las matrices FoldX 2R7G y 1N4M permiten clasificar péptidos como estables o inestables de acuerdo a la penalización que incluyen en cada posición. El conjunto de datos TP es limitado y consta de siete secuencias reportadas para Rb y cuatro secuencias reportadas para p107 en la base de datos de motivos lineales, ELM [36]. Por lo tanto, para SLiMs E2F contamos con pocas secuencias TP y las mismas tienen baja variabilidad de secuencia.

En el análisis realizado para las matrices 2R7G y 1N4M [33,38] (ver Figuras 7.8 y 7.9, Sección 7.6.4) se observó que ambas presentan limitaciones al penalizar residuos ubicados en la posición seis, que corresponde a la segunda posición fija del ***core*** del SLiM E2F definida por [NQDE] (Tablas 5.6 y 5.7). Esta posición si bien no se encuentra en contacto con los dominios *pocket*, forma parte de la hélice anfipática y determina la unión del SLiM E2F con el dominio [33,38]. Además, 1N4M admite la sustitución del residuo flanqueante al *core* del SLiM E2F, aspártico (D), en la posición nueve de la matriz (Tabla 5.7) por cualquier aminoácido, pero se sabe que este residuo acídico es importante para estabilizar la interacción del SLiM con Rb o p107 [33,38].

Por este motivo, se definieron variantes que acotan la matriz al ***core*** del SLiM con el fin de comparar cuál de ellas permite una mejor clasificación de los péptidos *hit* en estables o inestables. Para la matriz 2R7G se definieron tres variantes (Tabla 5.6): **2R7G_10** incluyendo las posiciones uno a diez (`PPTLHELYDL`); **2R7G_6** incluyendo las posiciones tres a ocho (`LHELYD`) y **2R7G_5** incluyendo las posiciones tres a siete (`LHELY`). Para la matriz 1N4M se definieron tres variantes (Tabla 5.7): **1N4M_9** abarcando los residuos uno a nueve (`GEGISDLFD`); **1N4M_6** que incluye los residuos cuatro a nueve (`ISDLFD`) y **1N4M_5** (`ISDLF`), que incluye residuos cuatro a ocho. En todos los casos, las variantes incluyen la secuencia ***core*** del SLiM E2F.

Tabla 5.6. Posiciones y residuos de la matriz 2R7G

	2R7G									
Posición#	1	2	3	4	5	6	7	8	9	10
Residuo	P	P	T	**L**	**H**	**E**	**L**	**Y**	D	L

#Diez posiciones de la matriz 2R7G [33], basada en la estructura cristalográfica del complejo formado por el péptido de la proteína viral E1A de adenovirus con el dominio *pocket* de Rb.
Rojo: C***ore*** del SLiM E2F.

Tabla 5.7. Posiciones y residuos de la matriz 1N4M

	1N4M								
Posición#	1	2	3	4	5	6	7	8	9
Residuo	G	E	G	**I**	**S**	**D**	**L**	**F**	D

#Nueve posiciones de la matriz 1N4M [38], basada en la estructura cristalográfica del complejo formado por el péptido de la proteína E2F2 humana con el dominio *pocket* de Rb.
Rojo: C***ore*** del SLiM E2F, presente en este péptido.

Con el fin de analizar la capacidad de las seis variantes de identificar a los TPs del SLiM E2F, se escanearon las secuencias de los interactores TP registrando la ocurrencia de la expresión regular del SLiM, y los mínimos de FoldX, marcando con un círculo amarillo las subsecuencias positivas para la REGEX y con un mínimo de FoldX (detección positiva) (Figura 5.5 y Figuras S6-S9, Anexo). Luego, se comparó el porcentaje de SLiMs TP detectados por cada variante de matriz (Tablas 5.8 y 5.9).

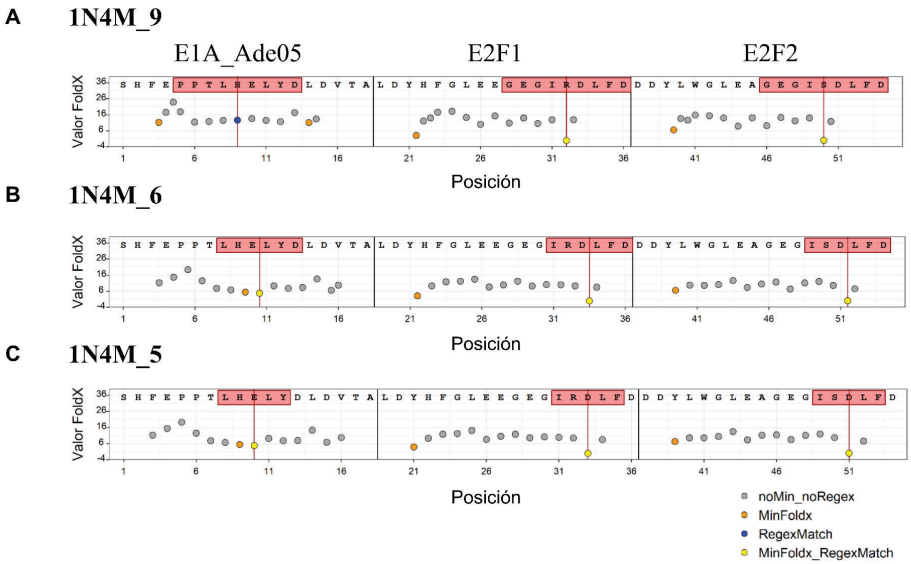


Figura 5.5. Distribución de valores FoldX de péptidos TP con SLiM E2F reportados para Rb. Fragmento de gráfico de puntos de péptidos TP con la proteína *pocket* Rb. Los 18 residuos de péptidos que pertenecen a las proteínas E1A de Adenovirus, E2F1 y E2F2 humanas, fueron escaneados con **A:** 1N4M_9, **B:** 1N4M_6 y **C:** 1N4M_5. El eje de abscisas indica la posición como el punto medio entre la posición inicial y final de cada subsecuencia analizada (abajo) y el detalle de residuo de la secuencia peptídica (arriba). Las líneas verticales rojas indican el punto medio de la subsecuencia donde se detectó la expresión regular `[IVLMA].[NQDE][IVLFMYAW][IVLFMYAW]`. Los puntos amarillos indican la detección de una subsecuencia conteniendo al SLiM E2F (recuadro rojo), con el valor mínimo de FoldX de todas las subsecuencias escaneadas para ese péptido. Los gráficos de puntos completos para Rb y p107 escaneados con las variantes de 2R7G y 1N4M se encuentran en las Figuras S6-S9, Anexo).

Tabla 5.8. Capacidad de detección en variantes de 2R7G en interactores conocidos.

Proteína pocket	TP (N-total)		2R7G_10*	2R7G_6**	2R7G_5***
Rb	7	% SLiMs detectados #	85	100	100
p107	4		100	100	100

% SLiMs detectados: % péptidos con valor mínimo de FoldX en la subsecuencia positiva para la *regex* del SLiM E2F.
*2R7G_10: variante de 2R7G de 10 residuos utilizada para el escaneo de secuencias.
**2R7G_6: variante de 2R7G de 6 residuos utilizada para el escaneo de secuencias.
***2R7G_5: variante de 2R7G de 5 residuos utilizada para el escaneo de secuencias.

Tabla 5.9. Capacidad de detección en variantes de 1N4M en interactores conocidos.

Proteína pocket	TP (N-total)		1N4M_9*	1N4M_6**	1N4M_5***
Rb	7	% SLiMs detectados #	85	100	100
p107	4		100	100	100

% SLiMs detectados: % péptidos con valor mínimo de FoldX en la subsecuencia positiva para la *regex* del SLiM E2F.
*1N4M_9: variante de 1N4Mde 9 residuos utilizada para el escaneo de secuencias.
**1N4M_6: variante de 1N4M de 6 residuos utilizada para el escaneo de secuencias.
***1N4M_5: variante de 1N4M de 5 residuos utilizada para el escaneo de secuencias.

Si bien el número de interactores conocidos (TP) es bajo para el SLiM E2F, las variantes más cortas de las matrices 2R7G y 1N4M tienen mejor *performance*, identificando todos los TP de Rb y p107 sin sacrificar especificidad (Tablas 5.8 y 5.9).

5.2.2. Análisis de recall y especificidad de interactores conocidos (TP) con SLiM E2F

Se utilizó el conjunto de datos TP y TN de Rb y p107 (Tabla S15 y S16) para el SLiM E2F para evaluar la capacidad de las matrices de distinguir entre ambos grupos. A partir de los TP, se crearon los TN mutando las posiciones fijas del SLiM a alanina (A). Se escanearon los TP y TN para Rb (N=7) y p107 (N=4) con las variantes de 2R7G y 1N4M determinando valores FoldX según el mismo procedimiento presentado para el SLiM LxCxE y se evaluaron los valores de *recall* y especificidad a diferentes valores umbral de FoldX (Figuras 5.6 y Figura Suplementaria S10, Anexo).

En el caso de la proteína Rb los valores FoldX de los interactores TP se encuentran solapados con los TN para las variantes 2R7G_10, 2R7G_6 y 1N4M_9, mientras que las variantes 2R7G_5, 1N4M_6 y 1N4M_5 fueron capaces de separar TP de TN (Figura 5.6). Para p107, cuyos TPs representan un subconjunto de los de Rb, todas las matrices logran separar TP de TN (Figura Suplementaria S10, Anexo). En todos los casos, en su conjunto los péptidos TP de ambas proteínas *pocket* presentan menores valores de FoldX que los péptidos TN (Tablas 5.10 y 5.11).

En el análisis de esta sección se observó que utilizando umbrales de cinco o seis se obtienen altos valores de *recall* y especificidad en todos los casos para las dos proteínas *pocket* (Figura 5.6 B y Figura Suplementaria S10, Anexo). Sin embargo, en el análisis de la sección 5.2.3, se observó que

considerando un umbral de tres, se obtienen generalmente porcentajes de *recall* más altos sin que la especificidad disminuya. Por este motivo, el análisis de este set de datos se realizó tomando como referencia un umbral de tres para evaluar la *performance* de las matrices FoldX 2R7G, 1N4M y sus variantes.

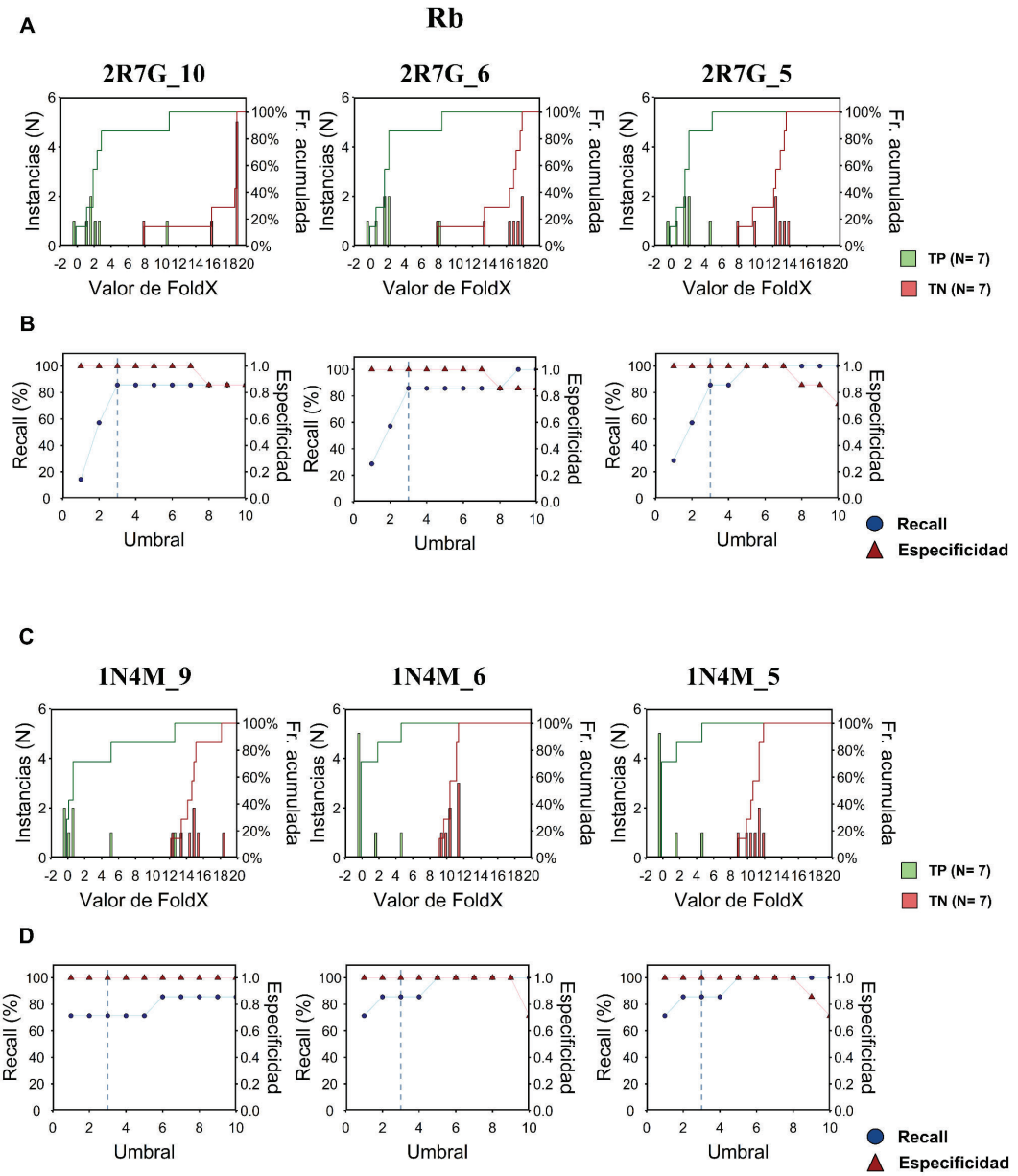


Figura 5.6. Interactores conocidos TP y TN de Rb escaneados con variantes 2R7G y 1N4M. Distribución de valores FoldX de péptidos TP y TN para matrices **A:** 2R7G y **C:** 1N4M. Eje Y izquierdo: Número de instancias, Eje Y derecho: porcentaje de frecuencia acumulada de instancias. **B:** R*ecall* (puntos azules) y especificidad (triángulos en rojo) de matrices 2R7G_10 (izquierda), 2R7G_6 (medio) y 2R7G_5 (derecha) para diferentes valores umbral de FoldX. **D:** Detalle del *recall* (puntos azules) y especificidad (triángulos rojos) de 1N4M_9 (izquierda), 1N4M_6 (medio) y 1N4M_5 (derecha). En ambos casos la línea vertical punteada señala un valor estimado de umbral de tres para establecer comparaciones.

Tabla 5.10. Métricas aplicadas a variantes de 2R7G en el escaneo de interactores conocidos.

Proteína Pocket	TP (N-total)	TN (N-total)	Métrica#	2R7G_10*	2R7G_6**	2R7G_5***
Rb	7	7	Recall (%)	86	86	86
			Especificidad	1	1	1
p107	4	4	Recall (%)	100	100	100
			Especificidad	1	1	1

Valores de *recall* y especificidad utilizando un punto de corte de tres.
*2R7G_10: variante de 2R7G de 10 residuos utilizada para el escaneo de secuencias.
**2R7G_6: variante de 2R7G de 6 residuos utilizada para el escaneo de secuencias.
***2R7G_5: variante de 2R7G de 5 residuos utilizada para el escaneo de secuencias.

Tabla 5.11. Métricas aplicadas a variantes de 1N4M en el escaneo de interactores conocidos.

Proteína Pocket	TP (N-total)	TN (N-total)	Métrica#	1N4M_9*	1N4M_6**	1N4M_5***
Rb	7	7	Recall (%)	75	86	86
			Especificidad	1	1	1
p107	4	4	Recall (%)	100	100	100
			Especificidad	1	1	1

Valores de *recall* y especificidad utilizando un punto de corte de tres.
*1N4M_9: variante de 1N4M de 9 residuos utilizada para el escaneo de secuencias.
**1N4M_6: variante de 1N4M de 6 residuos utilizada para el escaneo de secuencias.
***1N4M_5: variante de 1N4M de 5 residuos utilizada para el escaneo de secuencias.

Si bien el número de interactores TP E2F reportados para las proteínas *pocket* es bajo, los valores FoldX de los péptidos TP son menores que los péptidos TN, lo que sugiere que son más estables energéticamente. Sin embargo, en el caso de Rb, para las matrices 2R7G_10, 2R7G_6 y 1N4M_9 se observa que un TP posee valores iguales o superiores a péptidos TN. Las matrices restantes (2R7G_5, 1N4M_6 y 1N4M_5) poseen un desempeño similar.

5.2.3. Evaluación de péptidos *hit* de ProP-PD testeados experimentalmente con SLiM E2F

Con el objetivo de evaluar si las matrices FoldX son capaces de diferenciar la fuerza de unión de péptidos *hit* de ProP-PD, se escanearon las secuencias de los *hits* testeados experimentalmente utilizando las matrices 2R7G, 1N4M y sus variantes.

El análisis continuó con la distribución de péptidos *hit* de ProP-PD validados experimentalmente mediante ensayos *in vitro* en el laboratorio de trabajo, clasificados de acuerdo a la fuerza de interacción con la que se unen a las proteínas *pocket*, siendo la más fuerte SP (*strong positive binders*), débil W (*weak positive binders*) y sin interacción, N (*negative binders*). Estos péptidos cumplen con la expresión regular del SLiM E2F:

[IVLMA].[NQDE][IVLFMYAW][IVLFMYAW]

Se evaluó la estabilidad energética de 14 péptidos ensayados que fueron *hit* de Rb (Tabla S19, Anexo) y de 12 que fueron *hit* de p107 (Tabla S20, Anexo). Utilizando un umbral de tres, se calculó el porcentaje de *recall* y valor de especificidad para cada variante de matriz, considerando TP a la suma de *hits* SP +W y TN a los *hits* N (Figura 5.7 y Figura Suplementaria S11, Anexo).

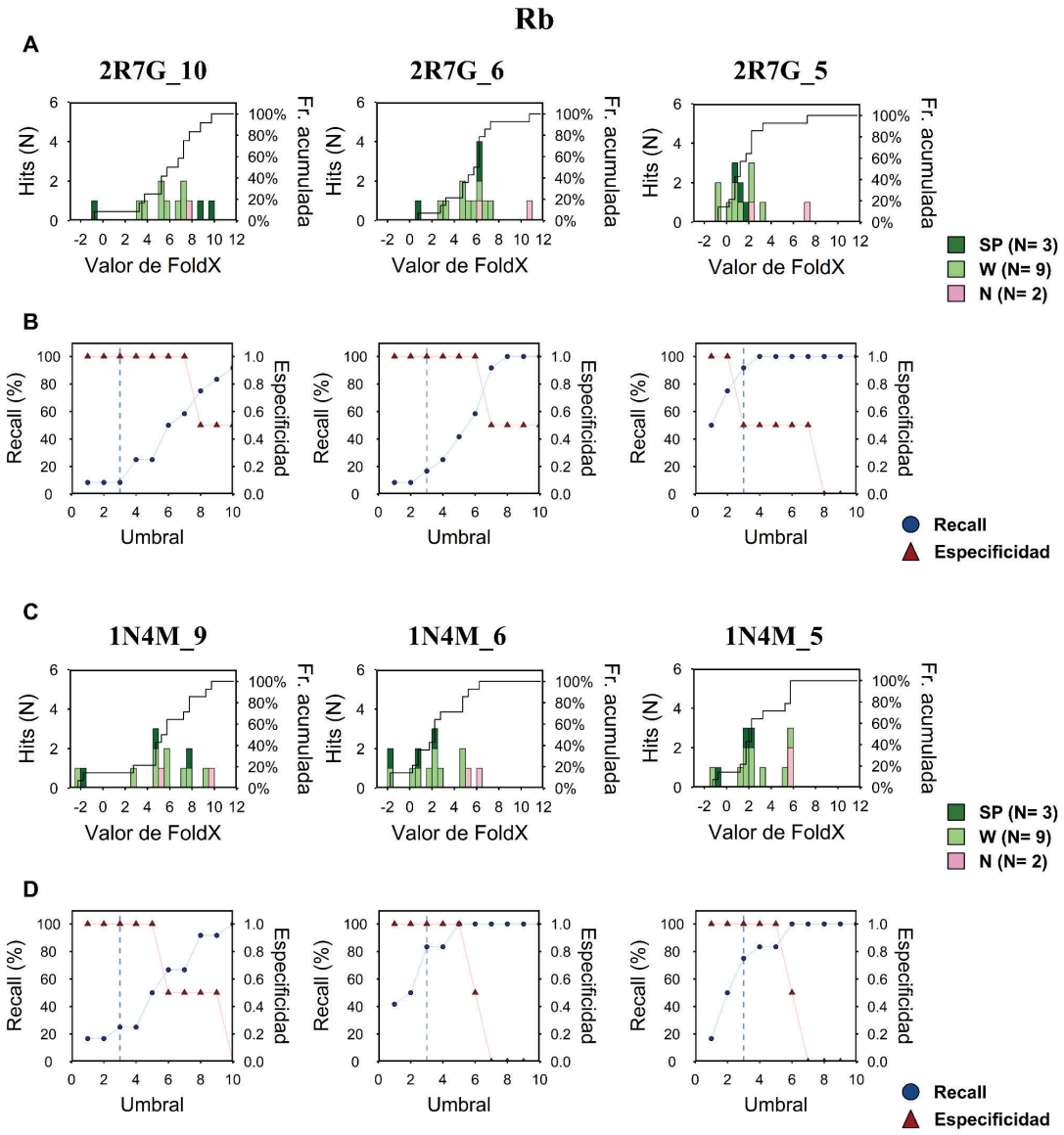


Figura 5.7. Péptidos *hit* en ProP-PD usando Rb como carnada que fueron testeados experimentalmente y escaneados con variantes de 2R7G y 1N4M. Distribución de péptidos testeados experimentalmente de acuerdo a valores FoldX de las variantes **A:** 2R7G y **C:** 1N4M. Eje Y izquierdo: Número de instancias, Eje Y derecho: porcentaje de frecuencia acumulada de instancias. **B:** R*ecall* (puntos azules) y especificidad (triángulos en rojo) de matrices 2R7G_10 (izquierda), 2R7G_6 (medio) y 2R7G_5 (derecha) para diferentes valores umbral de FoldX. **D:** Detalle del *recall* (puntos azules) y especificidad (triángulos rojos) de 1N4M_9 (izquierda), 1N4M_6 (medio) y 1N4M_5 (derecha). En ambos casos la línea vertical punteada señala un valor estimado de umbral de tres para establecer comparaciones.

Tabla 5.12. Métricas comparativas aplicadas al escaneo de péptidos *hit* de ProP-PD testeados experimentalmente, con las variantes de 2R7G.

Proteína Pocket	Binders SP+W (N-total)	Non Binders (N-total)	Métrica#	2R7G_10*	2R7G_6**	2R7G_5***
Rb	12	2	Recall (%)	8	16	92
			Especificidad	1	1	0,5
p107	5	7	Recall (%)	20	20	100
			Especificidad	1	1	0,3

Valores de *recall* y especificidad utilizando un punto de corte de tres.
*2R7G_10: variante de 2R7G de 10 residuos utilizada para el escaneo de secuencias.
**2R7G_6: variante de 2R7G de 6 residuos utilizada para el escaneo de secuencias.
***2R7G_5: variante de 2R7G de 5 residuos utilizada para el escaneo de secuencias.

Tabla 5.13. Métricas comparativas aplicadas al escaneo de péptidos *hit* de ProP-PD testeados experimentalmente, con las variantes de 1N4M.

Proteína Pocket	Binders SP+W (N-total)	Non Binders (N-total)	Métrica#	1N4M_9*	1N4M_6**	1N4M_5***
Rb	12	2	Recall (%)	25	83	75
			Especificidad	1	1	1
p107	5	7	Recall (%)	20	80	80
			Especificidad	0,9	0,6	0,6

Valores de *recall* y especificidad utilizando un punto de corte de tres.
*1N4M_9: variante de 1N4M de 9 residuos utilizada para el escaneo de secuencias.
**1N4M_6: variante de 1N4M de 6 residuos utilizada para el escaneo de secuencias.
***1N4M_5: variante de 1N4M de 5 residuos utilizada para el escaneo de secuencias.

Los mayores valores de *recall* y especificidad se logran manteniendo un valor de umbral de FoldX de tres considerado para el análisis de péptidos interactores o *binders* y no interactores o *non binders* para algunas de las variantes (Figura 5.7 B, D y Figura Suplementaria S11, Anexo).

Los resultados obtenidos para el análisis de *hits* ProP-PD validados experimentalmente, muestran que considerando un umbral de tres, las dos variantes con las que se obtiene al mismo tiempo un alto porcentaje de *recall* y mayor especificidad, son 1N4M_6 y 1N4M_5 (Tablas 5.12 y 5.13).

5.2.4. Evaluación de la *performance* de FoldX sobre todos los péptidos *hit* del ensayo ProP-PD

Para analizar si las matrices FoldX son capaces de distinguir si los péptidos que contienen al SLiM E2F son más estables que los que no presentan variantes del SLiM, se escanearon secuencias de estos dos grupos de péptidos y se evaluó su estabilidad energética utilizando las matrices 2R7G y 1N4M y sus variantes.

De los 308 péptidos *hits* de Rb, 173 no presentaban un SLiM detectado, 91 contenían un SLiM E2F y 12 incluían ambos SLiMs. En cuanto a p107, de los 103 *hits*, 40 no tenían un SLiM detectado, 8 contenían un SLiM E2F y 10 presentaban ambos SLiMs. En esta sección se analizarán los péptidos con SLiM E2F: **103 para Rb y 18 para p107**.

Empleando las variantes de 2R7G y 1N4M, se escanearon los péptidos que fueron *hit* en ProP-PD utilizando Rb y p107 como carnada. Los *hits* fueron agrupados según si se detectó en sus secuencia alguna variante del SLiM E2F o no:

1. Péptidos con variantes del SLiM E2F que cumplen con la expresión regular `[IVLMA].[NQDE][IVLFMYAW][IVLFMYAW]`,
2. Péptidos sin SLiM detectado.

Se utilizó el umbral establecido en secciones anteriores de tres para diferenciar péptidos estables de inestables entre los péptidos *hit* de ProP-PD utilizando la variantes de matrices 2R7G y 1N4M.

Se analizaron sub-secuencias escaneadas de ambos grupos, con el objetivo de identificar qué variante de las seis consideradas puede distinguir a los péptidos que contienen un SLiM E2F de los que no contienen al SLiM en términos de estabilidad energética (Figura 5.8 y Figura Suplementaria S12, Anexo).

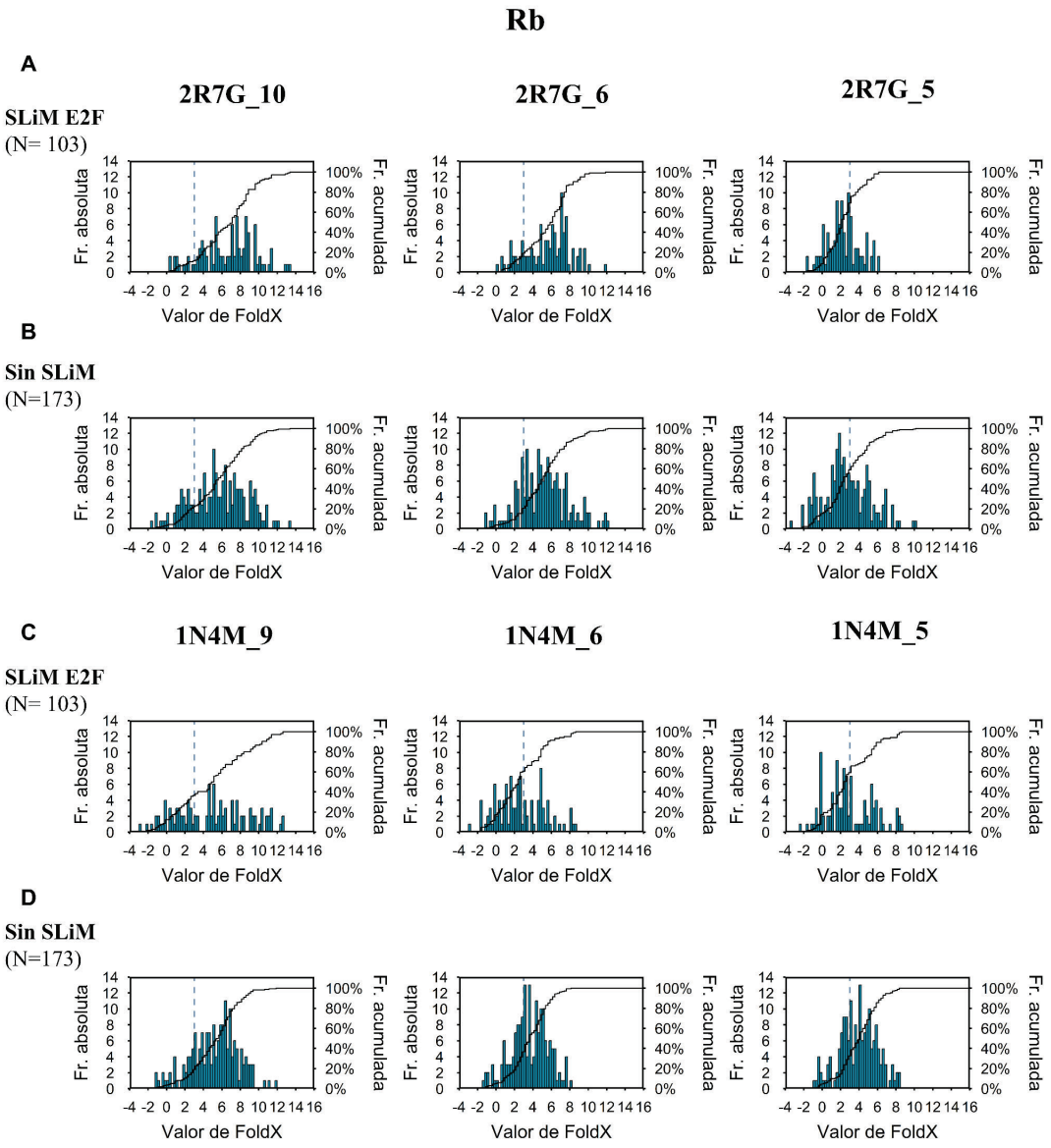


Figura 5.8 Péptidos *hit* de Rb escaneados con variantes de 2R7G y 1N4M. Distribución de valores FoldX de péptidos para matrices 2R7G_10, 1N4M_9 (izquierda), 2R7G_6, 1N4M_6 (medio), 2R7G_5 y 1N4M_6 (derecha). Eje Y izquierdo: Frecuencia absoluta de instancias, Eje Y derecho: porcentaje de frecuencia acumulada de instancias. La línea vertical punteada señala el valor de umbral utilizado para establecer comparaciones. **A:** Péptidos con SLiM E2F detectado escaneados con variantes de 2R7G. **B:** Péptidos sin SLiM E2F detectado escaneados con variantes de 2R7G. **C:** Péptidos con SLiM E2F detectado escaneados con variantes de 1N4M. **D:** Péptidos sin SLiM E2F detectado escaneados con variantes de 1N4M.

Se calculó como métricas el *recall* y la especificidad utilizando un umbral de tres para definir si alguna variante de 2R7G (Tabla 5.14) o 1N4M (Tabla 5.15) son capaces de distinguir ambos grupos. Se consideraron para estas métricas a los péptidos **con SLiM** como **interactores positivos (TP)** y a los péptidos **sin SLiM** como **interactores negativos (TN)**.

Tabla 5.14. Métricas comparativas aplicadas al escaneo de péptidos *hit* de ProP-PD con las variantes de 2R7G.

Proteína Pocket	TP (N-total)	TN (N-total)	Métrica #	2R7G_10*	2R7G_6**	2R7G_5***
Rb	103	173	Recall (%)	12	19	69
			Especificidad	0,8	0,8	0,4
p107	18	40	Recall (%)	17	17	61
			Especificidad	0,9	0,9	0,6

Valores de *recall* y especificidad utilizando un punto de corte de tres.
*2R7G_10: variante de 2R7G de 10 residuos utilizada para el escaneo de secuencias.
**2R7G_6: variante de 2R7G de 6 residuos utilizada para el escaneo de secuencias.
***2R7G_5: variante de 2R7G de 5 residuos utilizada para el escaneo de secuencias.

Tabla 5.15. Métricas comparativas aplicadas al escaneo de péptidos *hit* de ProP-PD con las variantes de 1N4M.

Proteína Pocket	TP (N-total)	TN (N-total)	Métrica #	1N4M_9*	1N4M_6**	1N4M_5***
Rb	103	173	Recall (%)	37	60	59
			Especificidad	0,8	0,6	0,7
p107	18	40	Recall (%)	11	50	50
			Especificidad	0,9	0,8	0,9

Valores de *recall* y especificidad utilizando un punto de corte de tres.
*1N4M_9: variante de 1N4M de 9 residuos utilizada para el escaneo de secuencias.
**1N4M_6: variante de 1N4M de 6 residuos utilizada para el escaneo de secuencias.
***1N4M_5: variante de 1N4M de 5 residuos utilizada para el escaneo de secuencias.

Utilizando un punto de corte de tres para establecer comparaciones entre las variantes de estas matrices FoldX, se observó que para el caso de 2R7G a mayores valores de *recall*, disminuye la especificidad (Tabla 5.14, 2R7G_5). Por el contrario, al utilizar las matrices 1N4M_6 y 1N4M_5 se obtiene un alto valor de *recall* y una especificidad cercana a uno (Tabla 5.15). Entre estas últimas dos, la que presenta un alto valor de *recall* y especificidad para ambas proteínas *pocket* es 1N4M_5, lo que sugiere que esta matriz presenta una mayor capacidad de distinción entre péptidos que contienen el SLiM E2F y péptidos que no.

Conclusiones generales de la *performance* de las matrices FoldX para el SLiM E2F: A partir del análisis realizado en los tres sets de datos con SLiM E2F se puede concluir que la variante 1N4M_5 es capaz de distinguir péptidos estables, considerados TP a aquellos que contienen SLiM E2F, de aquellos poco estables, que no contienen SLiM E2F considerados TN. Debido a que esta variante permite evaluar el *core* del SLiM, se utilizará para priorizar péptidos *hit* que contengan el SLiM E2F en ambas proteínas *pocket*. Los *hits* de ProP-PD serán considerados estables por debajo de un umbral de tres, lo que permitirá priorizar candidatos a futuros ensayos experimentales.

5.3. Relación entre estabilidad energética evaluada con FoldX y variantes de los SLiMs LxCxE y E2F presentes en los péptidos *hit*

El análisis de patrones de secuencia en péptidos *hit* se encuentra dentro de los parámetros de mayor importancia, ya que son determinantes para el éxito de la unión, según evidencia estructural y de ensayos de afinidad realizados entre las proteínas *pocket* e interactores [33,34,38,50].

Por otro lado, las matrices FoldX proporcionan información sobre la similitud entre la estabilidad energética del péptido *hit* y los péptidos en complejos de estructuras resueltas. Sin embargo, en algunas posiciones las penalidades de las matrices FoldX contradicen la evidencia experimental [33,34,38,50]. Por esta razón se optó por realizar el análisis con variantes de las matrices FoldX 1GUX y 1N4M que excluyen posiciones de conflicto, con el objetivo de acercarnos a lo que la evidencia experimental indica.

La evaluación conjunta de estos dos parámetros, presencia de una expresión regular (*regex*) conocida (ver Tablas 3.2 y 3.4, Secciones 3.2.1 y 3.2.2) y estabilidad energética, aporta un mayor grado de confianza permitiendo seleccionar *hits* que presentan variantes de las expresiones regulares de mayor afinidad y un valor de FoldX por debajo de los umbrales establecidos (Figura 5.9).

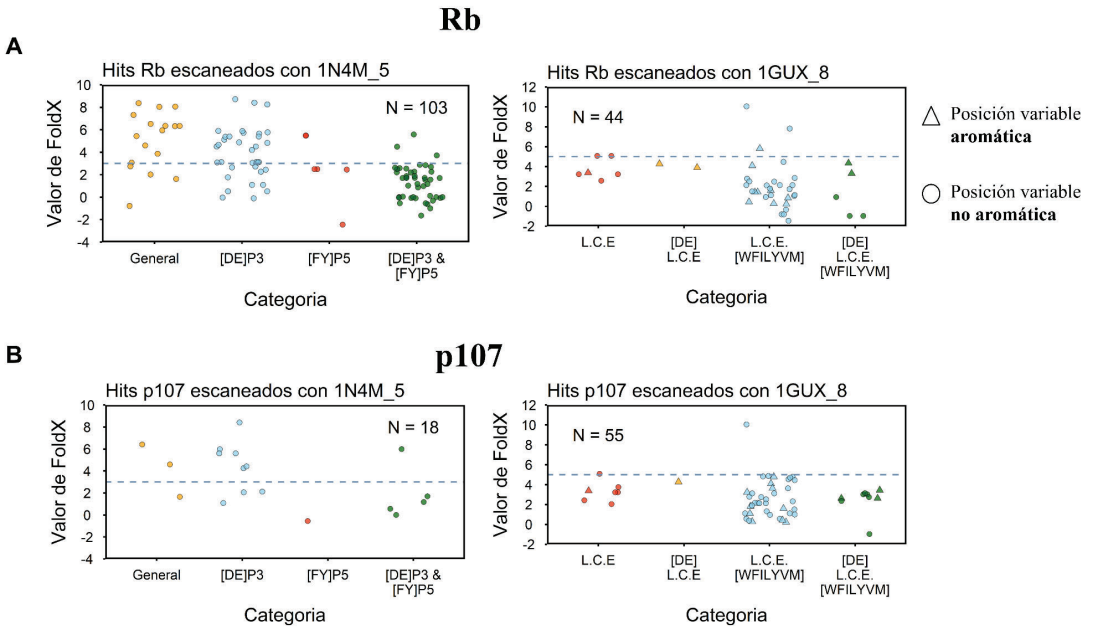


Figura 5.9. Valores de FoldX de péptidos *hit* de ProP-PD con diferentes variantes de REGEX. A: péptidos *hit* de Rb con variantes del SLiM E2F escaneados con 1N4M_5 (izquierda) y del SLiM LxCxE escaneados con 1GUX_8 (derecha). **B:** péptidos *hit* de p107 con variantes del SLiM E2F escaneados con 1N4M_5 (izquierda) y del SLiM LxCxE escaneados con 1GUX_8 (derecha). Los colores de los círculos indican una variante distinta del SLiM. Se indica el número total (N) de péptidos *hit* de Rb y p107 con la expresión regular detectada en los extremos superiores de cada gráfico. Triángulos: Presencia de un residuo F, Y o H en las posiciones variables del core del SLiM LxCxE.

El análisis de la Figura 5.9 revela que en la mayoría de los casos, los péptidos *hit* que contienen más de un residuo que modula favorablemente la afinidad de unión, poseen valores de FoldX por debajo de los umbrales definidos para cada matriz, y menores valores de FoldX con respecto a secuencias que solo contienen el ***core*** del SLiM (Figura 5.9). Esto sugiere que los patrones de secuencia deben considerarse al priorizar péptidos según su estabilidad energética.

En el análisis de péptidos con el SLiM LxCxE, se observó que aquellos que presentan un residuo hidrofóbico en las posiciones +2 o +3 respecto al *core* del SLiM, muestran valores de FoldX por debajo del umbral de cinco y en muchos casos por debajo de 2 (Figura 5.9, derecha). Esto respalda la inclusión de estas posiciones como fijas en el SLiM LxCxE al detectar patrones de secuencia (ver Sección 3.2.1). Por el contrario, la matriz 1GUX no puntúa con un menor valor de FoldX la presencia de residuos favorables en las posiciones variables en el *core* del SLiM LxCxE o la presencia de un residuo acídico en -1, mostrando que la matriz es limitada en la detección de estas características de alta afinidad [50].

En el caso de los péptidos con el SLiM E2F, se observó que aquellos que presentan los residuos [DE] en la segunda posición fija del SLiM y [FY] en la quinta posición, se encuentran mayormente por debajo del umbral de tres de FoldX, definido para este caso (Figura 5.9, izquierda). Aunque hay menos evidencia experimental disponible para los determinantes de unión del SLiM E2F en comparación con el SLiM LxCxE y los péptidos *hit* de p107 con el SLiM E2F fueron escasos, estos resultados contribuyen a priorizar con mayor confianza los péptidos con el SLiM E2F para su evaluación experimental, proporcionando así información valiosa sobre este SLiM.

5.4. Conclusiones generales sobre la determinación de la estabilidad energética de péptidos *hit* utilizando matrices FoldX.

Capacidad y limitaciones de matrices FoldX para representar la energética de unión de SLiMs: El análisis de estabilidad energética de los péptidos *hits* en ProP-PD utilizando matrices FoldX mostró que las posiciones flanqueantes del SLiM que modulan la afinidad de unión con los dominios *pocket*, no son adecuadamente representadas en las penalizaciones estimadas por FoldX. Esto ocurre por ejemplo, en las posiciones -1 (acídico, D) y +3 (hidrofóbico) con respecto al *core* del SLiM LxCxE, que se encuentran levemente penalizadas en la matriz 1GUX a pesar de ser importantes para modular la afinidad de unión [34,50]. Por lo tanto, las restricciones en estas posiciones no se encuentran representadas adecuadamente en 1GUX. Lo mismo ocurre con la posición nueve de la matriz 1N4M que corresponde a la posición flanqueante +1 con respecto al ***core*** del SLiM E2F. Esta posición, a pesar de ser un residuo acídico importante para estabilizar la unión, se encuentra poco penalizada por 1N4M [38].

Estas fallas en representar la energética de unión se deben en parte a que el algoritmo realiza

las sustituciones de manera independiente, sin alterar la orientación de la cadena principal del péptido. Como resultado, los residuos que no entran en contacto con la superficie del dominio, y que por lo tanto están menos restringidos en el espacio, reciben una penalización más baja. En cambio, las variaciones de energía libre suelen ser mayores cuando se sustituyen residuos que se entierran en el bolsillo de unión. Por lo tanto, FoldX sólo logra representar parcialmente la energética de interacción, y tampoco puede representar modos alternativos de unión que involucran cambios en la posición de la cadena principal.

Para mitigar el problema de las posiciones flanqueantes, se eliminó la posición -1 con respecto al *core* del SLiM LxCxE en la matriz 1GUX y en el caso de 1N4M se generó una variante que restringen la matriz al *core* del SLiM E2F, permitiendo limitar las penalizaciones a esos segmentos. Esta pérdida de representación de regiones flanqueantes, que modulan la afinidad de unión, puede ser compensada mediante la detección conjunta de variantes de expresiones regulares que incluyan estas posiciones y representan SLiMs de mayor afinidad.

Evaluación de la performance de diferentes variantes de matrices FoldX: A partir de la evaluación de sets de datos conteniendo el SLiM LxCxE utilizando las dos variantes de la matriz 1GUX, se observó que ambas son capaces de identificar péptidos más estables en interactores TP y que considerando un punto de corte de cinco, la variante 1GUX_8 permite obtener mayores valores de *recall* y especificidad. En el análisis de péptidos *hit* del ensayo ProP-PD que fueron testeados experimentalmente, se observó el mismo valor de *recall* y especificidad para ambas proteínas *pocket* y matrices, mientras que en el análisis de todos los péptidos *hit* se observó que tienen capacidad similar de distinguir entre péptidos que contienen el SLiM LxCxE de los que no lo presentan. Debido a que la matriz 1GUX_8 mostró un mejor desempeño en el escaneo de interactores conocidos, será empleada para realizar la priorización de péptidos candidatos a ensayos experimentales.

Para evaluar péptidos con el SLiM E2F se utilizaron las variantes de 2R7G y 1N4M. Las métricas establecidas mostraron que las matrices que están restringidas al *core* del SLiM presentan un mayor porcentaje de *recall* y alta especificidad en el escaneo de secuencias de interactores reportados (TP). Al analizar péptidos *hit* de ProP-PD testeados en el laboratorio, se pudo establecer un umbral de tres para distinguir entre péptidos estables e inestables. En este valor, se obtuvieron para las variantes cortas de 1N4M los valores más altos de *recall* y especificidad. Por último, al analizar todos los péptidos que fueron *hit* en ProP-PD se identificó a 1N4M_5 como la matriz con mayor capacidad de recuperación de péptidos estables que contienen el SLiM E2F sin sacrificar especificidad, siendo ésta superadora sobre las variantes de 2R7G y 1N4M restantes.

Comparativamente, el empleo de la matriz 1GUX para detectar péptidos con el SLiM LxCxE es superador con respecto al empleo de las variantes de 2R7G y 1N4M para detectar péptidos con SLiM E2F. Esto se observa en la distribución de valores que adoptan los péptidos con SLiM en comparación a los péptidos sin SLiM. Los péptidos en los que se identifica una variante de LxCxE,

presentan menores valores de FoldX que los que no presentan la *regex* (Figura 5.4 y Suplementaria S5, Anexo) al ser escaneados con cualquiera de las dos variantes de 1GUX. Por el contrario, este patrón se ve representado en el caso de péptidos conteniendo el SLiM E2F que fueron escaneados con 1N4M_5, siendo ésta una matriz acotada a las posiciones del *core* del SLiM, y no en el resto de las variantes 1N4M y 2R7G (Figura 5.8 y Suplementaria S12, Anexo). Por otro lado, el mayor número de interactores reportados con el SLiM LxCxE y la variabilidad de secuencia existente entre los mismos permite una mejor evaluación de las predicciones de la matriz 1GUX que las predicciones realizadas con las matrices 2R7G y 1N4M.

Conclusiones finales: Mediante el análisis de métricas de *recall* y especificidad de varias matrices sobre diferentes datasets de *benchmarking*, se logró seleccionar las matrices 1GUX_8 y 1N4M_5 para su uso posterior en la priorización de péptidos *hit*. A pesar de las limitaciones de las matrices FoldX, se concluye que la combinación de la detección de patrones de secuencia utilizando expresiones regulares con el análisis energético de los péptidos *hit* mediante FoldX permitirá generar una lista priorizada de péptidos, considerando valores menores de FoldX como un factor importante para la priorización, aunque no exclusivo.

Por último, como se mencionó al principio del capítulo, FoldX es una herramienta que permite predecir la afinidad de interacción de los complejos. Si bien en el set de datos de *hits* ProP-PD muchos de ellos fueron considerados negativos por no presentar la expresión regular definida de los SLiMs LxCxE y E2F, aquellos *hits* sin un SLiM conocido pero que tienen un valor bajo de FoldX, representan péptidos candidatos a interactuar de manera más afín con los dominios *pocket*. Estos *hits* podrían contener nuevas variantes de SLiMs de interacción que no han sido considerados en este trabajo o reportados en la bibliografía hasta el momento y podrían ser considerados en futuros ensayos experimentales.

Capítulo 6: Criterios de priorización para péptidos *hit*.

A lo largo de este trabajo se emplearon herramientas bioinformáticas que nos permitieron evaluar parámetros estructurales, de secuencia y de estabilidad energética de una lista de péptidos que fueron *hit* en el ensayo ProP-PD utilizando a los dominios *pocket* de Rb y p107 como carnada. Hacia el final del análisis, se consideró también la información de localización celular anotada por el grupo de investigación que diseñó la biblioteca HD2, siendo "núcleo" o "núcleo y citoplasma" las categorías de interés funcional dado que colocalizan con las proteínas *pocket* [40]. Se discutió la importancia de evaluar conjuntamente los resultados de los diferentes parámetros (secuencia, estructura y estabilidad) con el objeto de generar una lista priorizada de péptidos candidatos a ser testeados experimentalmente. El *benchmarking* de cada parámetro se realizó a partir de los valores obtenidos para una lista de péptidos que corresponden a SLiMs para los cuales se ha validado su interacción con las proteínas *pocket* curados a partir de la base de datos ELM y evidencias de nuestro grupo. Esto permitió establecer umbrales para identificar características deseables en los *hits*.

En primer lugar, se utilizaron expresiones regulares para priorizar péptidos conteniendo SLiMs candidatos en los que se detectaron patrones de secuencia que aumentan la afinidad de interacción con las proteínas *pocket*. En segundo lugar, se utilizaron parámetros estructurales para identificar péptidos que se encuentren con alta probabilidad accesibles para interactuar con las proteínas *pocket*. La implementación de RSA utilizando modelos AlphaFold permitió mejorar las predicciones de desorden utilizadas en el diseño de la biblioteca HD2 antes del advenimiento de este algoritmo de predicción estructural. Dentro de los péptidos con un valor de RSA mayor o igual a 0,4, se dió prioridad a los que presenten un mayor valor RSA y luego, a los que presenten un mayor valor de IUPred. Los que se identificaron con un valor de IUPred menor a 0,4 o bien compartían más de ocho residuos con dominios Pfam, fueron identificados con una advertencia. Aquellos que presentaron un valor de IUPred inferior a 0,2 y más de ocho residuos dentro de un dominio Pfam, fueron filtrados al igual que los que presentaron valores de RSA menores a 0,4. En tercer lugar, se priorizaron aquellos péptidos *hit* que se encontraban localizados en núcleo y núcleo-citoplasma, debido a que comparten localización celular con la familia de proteínas *pocket*. Por último, se evaluó que los SLiMs candidatos fuesen energéticamente estables según los valores predichos de estabilidad utilizando FoldX utilizando umbrales determinados mediante *benchmarking* con SLiMs conocidos.

En resumen, se detalla el orden en el que se aplicarán los criterios de priorización a lo largo de este capítulo:

1. **Detección de patrones de secuencia** que correspondan a las expresiones regulares definidas

para los SLiMs LxCxE y E2F. Los péptidos en los que no se detecten las expresiones regulares, serán filtrados.

2. **Evaluación de parámetros estructurales** que incluye:

a. **Valor de accesibilidad relativa al solvente (RSA):** se conservarán los péptidos que superen o igualen un valor de RSA de **0,4**. Los *hits* que se encuentren por debajo de este valor, serán filtrados.

b. **Valor de predicción del desorden con IUPred:** se priorizarán péptidos con un promedio de valor de IUPred superior o igual a 0,4. Aquellos que se encuentren con un valor menor a **0,4** serán identificados con una advertencia.

c. **Solapamiento con residuos Pfam:** se priorizarán péptidos que compartan ocho residuos o menos con dominios Pfam. Aquellos péptidos que compartan nueve o más, serán identificados con una advertencia.

d. Si el péptido supera el valor de 0,4 de RSA, pero presenta un valor de IUPred inferior a **0,2** y más de **ocho** residuos solapados con un dominio Pfam, será filtrado.

3. **Localización celular:** Se priorizarán péptidos con localización celular nuclear y nuclear-citoplasmática.

4. **Estabilidad energética según matrices FoldX:**

a. Péptidos con SLiM LxCxE escaneados con 1GUX_8: se identificarán con advertencias los que se encuentren por encima de un valor de cinco.

b. Péptidos con SLiM E2F escaneados con 1N4M_5: se identificaran con advertencias los que se encuentren por encima de un valor de tres.

Por último, para los *hits* que cumplen con los criterios estructurales y que poseen localización citoplasmática, serán considerados en la priorización dependiendo de la función asociada. Esto se debe a que algunas proteínas que podrían mediar los roles mitóticos de las proteínas *pocket,* tienen asignada localización citosólica debido a que la membrana nuclear se disgrega durante la mitosis. En este trabajo, esta selección fue realizada para algunos *hits* mediante curación manual, pero en el futuro podría automatizarse mediante la incorporación de términos GO en la priorización. Aquellos *hits* provenientes de proteínas reportadas en IntAct o ensayos de proteómica en los que no se haya detectado una expresión regular previamente conocida, serán considerados ya que podrían representar modos de interacción novedosos. Estas consideraciones guiarán la elección de candidatos priorizables discutidos a lo largo de este capítulo.

6.1. Ejemplos de uso de criterios de priorización en péptidos *hit* LxCxE de Rb testeados experimentalmente

A continuación, a modo de ejemplo y con el objetivo de explicar la relevancia de los distintos

criterios se aplican los mismos al conjunto de péptidos con el SLiM LxCxE identificados como *hits* de Rb y previamente validados en el laboratorio. De los 26 péptidos con SLiM LxCxE ensayados (Tabla S17, Anexo), este análisis incluye 11 péptidos *hit* que provienen de la biblioteca HD2 (Tabla 6.1).

Tabla 6.1. Criterios de priorización aplicados a Péptidos *hit* de Rb testeados experimentalmente con SLiM LxCxE.

ID Uniprot	RSA[#]	IUPred[#]	Pfam[#*]	Localización celular[# †]	FoldX[#**]	Tipo de interacción***	Criterio aplicado
CPSF7_HUMAN	0,83	0,46	0	N	4,34	SP	Priorizable
FANCM_HUMAN	0,83	0,37	0	N	1,78	W	Priorizable (IUPred)
HFM1_HUMAN	0,88	0,20	0	NC	1,43	W	Priorizable (IUPred)
HMBX1_HUMAN	0,52	0,65	16	NC	0,43	N	Priorizable (Pfam)
GTSE1_HUMAN	0,75	0,33	15	NC	1,50	SP	Priorizable (IUPred y Pfam)
SEPT7_HUMAN	0,46	0,26	16	NC	4,27	SP	Priorizable (IUPred y Pfam)
S31D1_HUMAN	0,73	0,41	16	C	2,13	SP	Priorizable (Pfam, Localización)
KIF24_HUMAN	0,81	0,62	0	C	0,22	SP	Priorizable (Localización)
CE295_HUMAN	0,76	0,51	0	C	3,38	SP	Priorizable (Localización)
ZN445_HUMAN	0,62	0,51	0	N	10,05	W	Priorizable (FoldX)
AF17_HUMAN	0,48	0,10	16	N	1,72	SP	Filtrado (IUPred, Pfam)

[#]Celdas en amarillo indican una advertencia. Celdas en rojo señalan el motivo de filtrado
* Número de residuos solapados con dominios Pfam
[†]N= Núcleo; NC= Núcleo y Citoplasma; C= Citoplasma
** Valor de FoldX con la matriz FoldX 1GUX_8
*** Tipo de interacción con las que fueron identificadas en ensayos experimentales con el dominio *pocket* de Rb, siendo SP: *Strong Positive binder*, W: *Weak positive binder* y N: *Negative binder*.

Todos los *hits* cumplen con el primer criterio (presencia de expresión regular), dado que presentan en su secuencia una variante del SLiM LxCxE. El segundo criterio, incluye a los tres parámetros estructurales considerados en este trabajo (RSA, IUPred y Pfam).

Luego, se analizaron según la localización celular anotada para cada péptido en la biblioteca HD2, siendo núcleo o núcleo y citoplasma las de preferencia y por último que al escanearlos con la matriz FoldX 1GUX_8 tengan un valor menor a cinco (energéticamente estables). Los 11 péptidos fueron clasificados según estos criterios en 'priorizable', 'priorizable con advertencia', 'no priorizable' y 'filtrado'; y se analizaron algunos casos de manera individual.

Dentro de este conjunto de péptidos *hits* evaluados experimentalmente se observan siete casos interesantes que resaltan la relevancia de los criterios.

Importancia del valor de RSA. El primer caso es la proteína HFM1_HUMAN, con un valor de IUPred de 0,2. La proteína HFM1 es una DNA helicasa ATP dependiente. Este péptido se encuentra en el N-terminal de la proteína que en el modelo AlphaFold es desordenado y está accesible para interactuar (RSA= 0,88). No existe evidencia alguna hasta la fecha que reporte una estructura

ordenada. El bajo valor de IUPred puede deberse a un efecto borde del algoritmo, que tiende a predecir valores más bajos en los extremos de las secuencias ya que se usa una ventana de residuos más chica para predecir el desorden. Este ejemplo resalta la importancia de complementar las predicciones de IUPred con herramientas como RSA, que nos acerquen a identificar candidatos para validaciones experimentales.

Importancia de residuos solapados con dominios Pfam:

Dominios Pfam y Dimerización: El segundo caso es el péptido de la proteína HMBX1_HUMAN. HMBX1, Proteína con homeobox 1, es un factor de transcripción que pertenece a la clase HNF de la familia de genes homeobox. Los valores de FoldX, RSA e IUPred cumplen con los criterios de priorización y se detecta la expresión del SLiM LxCxE en su secuencia (Tabla S17, Anexo). Sin embargo, el péptido se encuentra en el N-terminal solapado con un un dominio Pfam denominado HNF1_N, cuya función es de dimerización [63]. Dado que la base de datos de AlphaFold2 sólo contiene modelos monoméricos, el RSA calculado a partir del modelo de esta proteína podría reflejar incorrectamente la accesibilidad de este péptido para interactuar en el caso de que esté formando el dímero [54]. Por lo tanto, si bien se determinó experimentalmente que no interactúa con Rb, si fuera positivo podría haber sido un falso positivo requiriendo otros experimentos que garanticen su accesibilidad.

Dominios Pfam y Regiones Desordenadas: El péptido proveniente de GTSE1_HUMAN, la proteína 1 de expresión en fase G2 y S, solapa con el dominio Pfam GTSE1_N. El péptido *hit* de la S31D1_HUMAN, proteína 31D1 espermatogénesis asociada, se encuentra dentro de un dominio Pfam DUF4599, que son dominios sin función conocida. Una búsqueda bibliográfica no reveló información sobre la estructura de estos dos dominios y los modelos AlphaFold indican que ambas proteínas son desordenadas en esa región, por lo que puede tratarse de regiones desordenadas (IDRs) de alta conservación que se clasifican como dominios Pfam sin ser globulares y por ende no deben ser descartadas.

Dominios Pfam y Loops Desordenados: El péptido de la proteína septina-7 SEPT7_HUMAN presenta valores cercanos a los umbrales de FoldX, RSA e IUPred. El péptido está completamente solapado con un dominio septina de Pfam. Sin embargo, existen 15 cadenas que representan el dominio provenientes de seis estructuras determinadas por cristalografía de rayos X (PDBs: 2qag, 3tw4, 6n0b, 6n12, 6uqq, 3t5d) y una determinada por microscopía electrónica (PDB: 7m6j). En una de estas cadenas, la región correspondiente al péptido *hit* se encuentra como *missing residues*, es decir, no se pudo resolver la estructura, mientras que en las restantes es un *loop* intradominio de estructura variable (Figura 6.1A). Esto indica que el péptido está expuesto y accesible para interactuar con el dominio globular más allá de la advertencia de Pfam.

La detección de dominios Pfam de manera complementaria a las otras herramientas permite detectar aquellos *hits* que podrían resultar en falsos positivos por encontrarse dentro de dominios

globulares o no accesibles, así como también indicar para cuales *hits* es necesario profundizar en la búsqueda bibliográfica como el caso de SEPT7_HUMAN. Los *hits* con **advertencias** de Pfam deben ser analizados de manera individual y teniendo en cuenta el resto de los parámetros evaluados.

Utilización de FoldX para la priorización de SLiMs alternativos. SEPT7 también permite ejemplificar el uso de FoldX para seleccionar el mejor SLiM candidato. En la secuencia de SEPT7 se detectaron patrones de los SLiMs E2F (subrayado) y LxSxE (indicado en azul):

SLF**LTDLYSPE**YPGPS

Al ser escaneado con la matriz LxCxE (1GUX_8), el péptido tiene un valor de FoldX de 4,27, cercano al umbral de cinco utilizado para este SLiM, mientras que con la matriz E2F (1N4M_5) obtiene un valor de 1,70, cercano a la mitad del umbral utilizado para este SLiM. Esto sugiere que el péptido de SEPT7 podría unirse preferencialmente al bolsillo E2F, en lugar del LxCxE. De hecho, el ensayo de competencia Alphascreen realizado en el laboratorio sugiere que este péptido se une al bolsillo E2F y no al bolsillo LxCxE. El laboratorio está llevando a cabo ensayos experimentales de mutagénesis para confirmar esta hipótesis.

Limitaciones de FoldX en la detección de SLiMs. El péptido de la proteína Zinc Finger 445 (ZN445_HUMAN) presenta valores de parámetros estructurales que indican que podría ser priorizado y una localización celular deseada, pero posee un valor de FoldX de 10,05 (**A**R**EPW**). Este valor de FoldX se debe a que en la posición hidrofóbica +2 hay un triptófano que es altamente penalizado por FoldX. Sin embargo, este péptido fue un interactor positivo en los ensayos experimentales y nuestro grupo determinó que la mutación de leucina a triptófano en la posición +2 equivalente del SLiM LxCxE de HPV E7 es poco disruptiva, disminuyendo la afinidad en sólo 2 veces (ver Tabla S2, Anexo) indicando que el triptófano es una variante posible y que la penalización extrema del triptófano en la posición +2 es una limitación de FoldX.

Concluímos que el empleo de matrices FoldX (ver Secciones 5.1 y 5.2) para identificar péptidos estables es complementaria al resto de las herramientas ayudando en la identificación de SLiMs, pero no debe ser excluyente debido a las limitaciones del algoritmo.

Relevancia de IUPred y un filtrado exitoso. El último caso de interés es el péptido de la proteína AF17_HUMAN, identificado como SP en ensayos experimentales, y que presenta un residuo hidrofóbico en la posición +2, lo que le confiere una mayor afinidad de unión y los valores de FoldX y RSA son apropiados. Sin embargo, el valor de IUPred predice que la secuencia es ordenada, con un valor de 0,1. El modelo AlphaFold muestra que forma parte de una lámina beta dentro de un dominio zinc-finger globular al igual que ocurre en la proteína homóloga AF10 (Figura 6.1 B). Este caso resulta ser un falso positivo ya que aunque el péptido aislado es un interactor fuerte, en condiciones

nativas se encuentra inaccesible. Por lo tanto, la evaluación de péptidos que presentan un valor de IUPred por debajo de 0,2 en conjunto con una advertencia de Pfam, podría facilitar el filtrado de péptidos que no son candidatos priorizables por encontrarse dentro de un dominio globular, aún cuando el valor de RSA es adecuado (Figura 6.1 B).

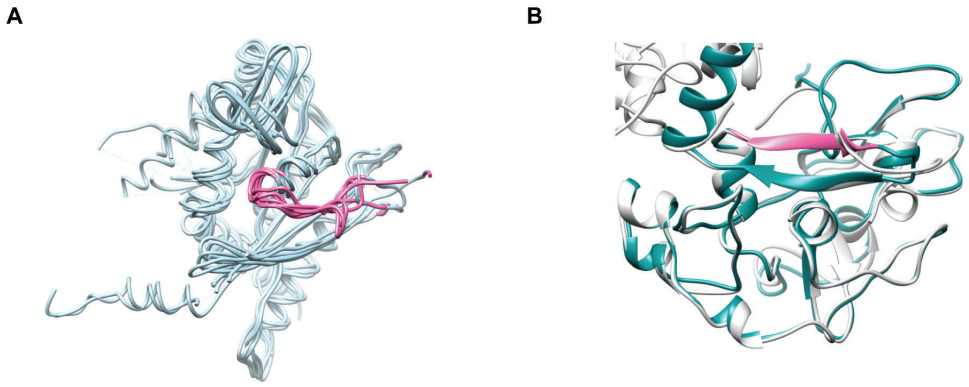


Figura 6.1. Estructuras de segmentos proteicos conteniendo péptidos hit de Rb, testeados experimentalmente. A: Estructura alienada de los PDB: 2QAG, 3TW4, 6N0B, 6N12, 6UQQ, 3T5D y 7M6J [64] de la proteína SEPT7. Se observa la estructura en color azul claro y se destaca el segmento correspondiente al péptido *hit* en color rosa. **B:** Estructura modelada en AlphaFold2 (AF-P55198-F1 [23]) de la proteína AF17 (verde) alineada con la estructura cristalizada del dominio PZP de la proteína AF10 (gris) (PDB: 5DAH) [64] donde se detectó el patrón del SLiM LxCxE con un residuo hidrofóbico en la posición +2 respecto al core del SLiM (**L**L**C**E**E**E**V**) que forma parte de una lámina beta dentro de un dominio zinc-finger.

Estos siete ejemplos demuestran la importancia de realizar un análisis integral que incluya más de una herramienta y umbrales que permitan analizar parámetros con advertencias.

6.2. Mapeo de sitios de unión en interactores conocidos (IntAct y proteómica)

En esta sección, se priorizarán aquellos péptidos *hit* de ProP-PD que son interactores previamente conocidos de las proteínas *pocket* reportados en IntAct [42] (ver Sección 2.2) o en ensayos de proteómica [48] (ver Sección 2.3).

i- Candidatos provenientes de la base de datos IntAct: Se identificaron tres *hits* de Rb, dos sin SLiM conocido (ECD_HUMAN, XPA_HUMAN) y uno con un SLiM LxAxE (TAF1_HUMAN) (Tabla 6.2).

El *hit* de TAF1_HUMAN, se solapa en once residuos con el dominio Pfam de unión a la TATA-binding-protein (TBP-binding). Sin embargo, los 300 residuos N-terminales de TAF1 son asignados como *missing residues* en 24 estructuras obtenidas por microscopía electrónica excepto en una estructura donde se observan 60 residuos de esta región (Figura 6.2). Esto lo convierte en un *hit* a **priorizar** por encontrarse el SLiM candidato en una región desordenada.

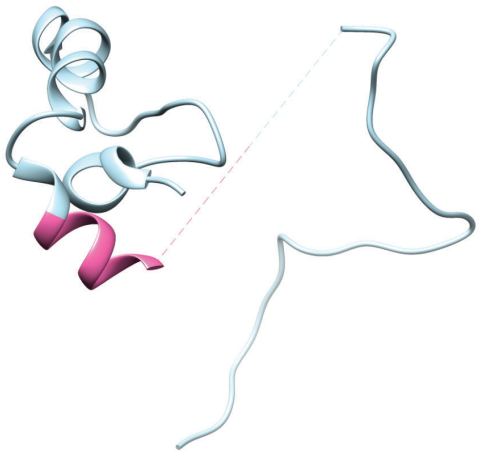


Figura 6.2. Modelo de estructura de TAF1. Se observa la estructura de la cadena A del PDB: 6MZD [64] correspondiente a la proteína TAF1 humana. Se indica la estructura en color azul claro y la región del péptido *hit* en color rosa. Los seis primeros residuos forman una estructura de hélice, mientras que los otros diez residuos figuran como *missing residues*, con línea rosa punteada.

Tabla 6.2. Péptidos *hit* de ProP-PD reportados en IntAct y en ensayos de proteómica

ID Uniprot[#]	Secuencia[##]	SLiM[Δ]	RSA	IUPred[###]	Pfam* [###]	Localización [†]	FoldX 1GUX_8**	FoldX 1N4M_5***	Criterio aplicado
TAF1_HUMAN **(I)**	SLITE**LTANEELT**GTD	No reportado LxAxE	0.77	0.50	11	N	0,93	-	Priorizable (Pfam)
ECD_HUMAN **(I)**	SVMAPVDVDLNLVSNI	-	0,47	0,58	9	NC	4,21	3,87	No priorizable (*Regex*)
XPA_HUMAN **(I)**	LEVWGSQEALEEAKEV	-	0,42	0,42	0	NC	5,12	6,61	No priorizable (*Regex*)
TPM3_HUMAN **(P)**	**MMEAI**KKKMQMLKLDK	E2F	0,68	0,48	0	C	-	3,05	No priorizable (FoldX)
DREB_HUMAN **(P)**	S**LIDLW**PGNGEGASTL	E2F	0,88	0,48	0	C	-	4,83	Priorizable (FoldX)
PRUN2_HUMAN **(P)**	SG**IMELY**GSDIEPQPS	E2F	0,82	0,53	0	C	-	-1	Priorizable (Localización)
	PT**FLEIW**NDSVDGDSF	E2F *like*	0,87	0,48	0	C	8,0	5,89	Priorizable (FoldX)
	DRKTPT**FLEIW**NDSVD		0,86	0,45	0		8,15		Priorizable (FoldX)
PDLI1_HUMAN **(P)**	VTEEGKRHPYKMNLAS	-	0,77	0,60	0	C	7,82	4,16	Priorizable (FoldX)
PLAK2_HUMAN **(P)**	LTVVKDDDHGILDQFS	-	0,52	0,81	0	NC	7,36	-0,41	Priorizable (*Regex*)
SCMC1_HUMAN **(P)**	ELLKSYW**LDNFA**KDSV	E2F	0,42	0,03	11	C	-	5,43	Filtrado (IUPred, Pfam)

[#]Se indica entre paréntesis si el péptido fue reportado en IntAct **(I)** o en ensayos de proteómica **(P)**
[##]Detección de la expresión regular de los SLiMs LxCxE o E2F (negro), detección de una variante similar (rojo). Subrayados se encuentra la secuencia que corresponde al menor valor de FoldX de los reportados en esta tabla.
[###]Celdas en amarillo indican una advertencia. Celdas en rojo señalan el motivo de filtrado
[Δ]SLiM detectado.
* Número de residuos solapados con dominios Pfam
[†]N= Núcleo; NC= Núcleo y Citoplasma, C= Citoplasma
**Valor mínimo de secuencia escaneada con matriz FoldX 1GUX_8.
***Valor mínimo de secuencia escaneada con matriz FoldX 1N4M_5.

Los *hits* de las proteínas ECD y XPA cumplen con los criterios estructurales propuestos en este trabajo y ambos son identificados con localización núcleo-citoplasmática. Sin embargo, en el

modelo AlphaFold de ECD se observa que nueve residuos del péptido forman parte de una hélice estructurada. En el caso de XPA en once cadenas de diez estructuras y en el modelo AlphaFold, los 16 residuos forman parte de una hélice estructurada. Por lo tanto, estos péptidos no serán priorizados.

ii- Candidatos provenientes del ensayo de Proteómica: Se identificaron siete péptidos no solapados pertenecientes a seis proteínas del ensayo de proteómica [48] (Tabla 6.2). Dos de ellos no presentaron los patrones de secuencia correspondientes a los SLiMs E2F y LxCxE (PDLI1_HUMAN y PLAK2_HUMAN), y cinco pertenecen a cuatro proteínas clasificadas en la biblioteca HD2 como citoplasmáticas (SCMC1_HUMAN, TPM3_HUMAN, DREB_HUMAN, PRUN2_HUMAN).

Entre las proteínas de localización citosólica, la proteína mitocondrial **SCMC1_HUMAN**, posee un valor muy bajo de IUPred (0.03) y se encuentra solapado con el dominio Pfam "*Mitochondrial carrier protein*". Si bien no existe una estructura resuelta, en el modelo AlphaFold2 se encuentra formando una hélice indicando que el SLiM presente en su secuencia no se encontraría accesible. La proteína **TPM3_HUMAN**, posee funciones no relacionadas con las proteínas *pocket* por lo que no será priorizada. Por el otro lado, el péptido de **DREB_HUMAN**, se identificó un SLiM E2F, esta proteína citosólica posee roles en diferenciación celular por lo que sí será priorizada.

Por otro lado, se identificaron dos péptidos no solapados de la proteína citoplasmática **PRUN2_HUMAN**: uno de ellos presentó un SLiM E2F canónico (`IMELY`) que fue ensayado experimentalmente en el grupo de trabajo resultando un interactor débil, y el otro péptido presenta una variante novedosa del SLiM E2F (*E2F-like*) con una fenilalanina en la primera posición del SLiM (**F**`LEIW`) que resta testear y presenta buenos valores estructurales (Tabla 6.2). Si bien la localización de PRUN2_HUMAN es citosólica, es priorizable por estar relacionada al desarrollo tumoral y por su expresión en la fase G1 del ciclo celular. Estos péptidos están siendo testeados en el laboratorio.

La proteína **PLAK2_HUMAN** por otro lado, presenta parámetros estructurales priorizables y localización nuclear (Tabla 6.2). Aunque no presenta en su secuencia una expresión regular canónica para los SLiMs E2F y/o LxCxE, podría representar una variante *E2F-Like* `ILD`**`Q`**`F` de interacción con el dominio *pocket* ya que esa secuencia puntúa muy bien con la matriz (FoldX 1N4M_5: -0,41), y en otros *hits* sin SLiM detectado como KNL1 testeados en el laboratorio se detectó un SLiM funcional `LTD`**`T`**`W` donde la segunda posición hidrofóbica también es reemplazada por un residuo polar.

En resumen, se identificaron los siguientes péptidos candidatos a priorizar a TAF1_HUMAN (SLiM LxAxE), PRUN2_HUMAN (SLiMs E2F/E2F-Like) y PLAK2_HUMAN (SLiM E2F-Like), DREB_HUMAN (E2F). PRUN2_HUMAN y DREB_HUMAN resaltan la relevancia de un futuro profundizar el análisis considerando términos GO de *hits* cuyos roles se encuentren relacionados a los de las proteínas *pocket*.

6.3. Variantes novedosas del SLiM E2F identificadas con MEME.

En esta sección, se priorizaron los candidatos que surgieron del alineamiento con la herramienta MEME [52] en los que se identificaron variantes novedosas del SLiM E2F (*E2F-like*) incluyendo un SLiM E2F invertido (E2F reverso) (ver Sección 3.3). Los modos de unión "reverso" ya han sido validados para otros SLiMs de la base de datos ELM.

Se identificaron péptidos solapados provenientes de dos proteínas con una fenilalanina en la primera posición del SLiM E2F propuesto en este trabajo. Estos *hits*, provienen de las proteínas PRUN2_HUMAN y ZN865_HUMAN. En el mismo análisis, MEME [52] incluyó un alineamiento con la secuencia del péptido de la proteína estriatina-1 (STRN3_HUMAN) donde no se detectó similitud a ningún SLiM conocido (Tabla 6.3).

Tabla 6.3. Péptidos *hit* de ProP-PD con E2F-like y E2F reverso.

ID Uniprot	Secuencia[#]	SLiM*	RSA	IUPred	Pfam	Localización[†]	FoldX 1GUX_8**	FoldX 1N4M_5***	Criterio aplicado
ZN865_HUMAN	SYPFDFLEFLNHQRFE	E2F *like*	0,83	0,46	0	N	6,13	4,1	Priorizable (*Regex*)
	VHFQSYPFDFLEFLNH		0,83	0,46	0		7,64		Priorizable (*Regex*)
STRN3_HUMAN	VLETFNFLENADDSDE	E2F reverso	0,89	0,64	0	NC	5,03	0,13	Priorizable (*Regex*)

[#]Detección de la expresión regular de los SLiM *E2F-like* (rojo), E2F reverso (verde) o secuencia alineada por la herramienta MEME (azul). Subrayado se indica la secuencia que corresponde al menor valor de FoldX reportado en la tabla
*SLiM detectado.
[†]N= Núcleo; NC= Núcleo y Citoplasma
**Valor mínimo de secuencia escaneada con matriz FoldX 1GUX_8.
***Valor mínimo de secuencia escaneada con matriz FoldX 1N4M_5.

Los péptidos provenientes de **PRUN2_HUMAN** fueron discutidos en la sección anterior, dado que la proteína fue identificada como interactora de Rb en ensayos de proteómica [48]. Los péptidos provenientes de la proteína **ZN865_HUMAN** cumplen con los criterios estructurales y se localizan en núcleo. El valor mínimo de FoldX con 1N4M_5 corresponde a un péptido con prolina, sin embargo, el SLiM E2F forma una hélice, por lo que una prolina no es aceptable. El valor FoldX correspondiente a la secuencia FLEFL es de 5,4 superior al umbral de tres elegido, lo que resulta de la sustitución por fenilalanina en la primera posición (Tabla 6.3 y Tabla S12, Anexo). Pero podría ser que FLEFL sea la secuencia de unión y la matriz FoldX no refleje bien la aceptación de F en la primera posición. En el péptido de **STRN3_HUMAN** se identificó un SLiM E2F reverso, que podría representar una forma novedosa de interacción con el dominio *pocket,* dado que requiere el posicionamiento invertido de la hélice que forma el SLiM E2F al unirse con Rb. Este péptido presenta parámetros estructurales y de localización celular priorizables. El valor de FoldX no corresponde con la secuencia propuesta, pero de ser un E2F reverso no puede considerarse ya que la orientación del péptido sería la inversa (Tabla 6.3). Considerando el valor de FoldX, también podría ser un caso similar al de KNL1_HUMAN LTD**T**W discutido en la sección 6.2 donde la segunda posición hidrofóbica es reemplazada por un residuo polar, en este caso la secuencia FLE**N**A posee un valor de

FoldX 8,8 que resulta de la sustitución por fenilalanina en la primera posición y alanina en la última (Tabla S12, Anexo). La secuencia detectada por FoldX VLE**T**F también es un caso similar a KNL1_HUMAN. Sólo la evaluación experimental podrá desambiguar estos casos.

Los tres *hits* discutidos en esta sección, representan variantes novedosas del SLiM E2F (E2F-Like y E2F-Reverso) que han sido detectados gracias al análisis bioinformático presentado y constituyen ejemplos de alto interés para ser evaluados experimentalmente.

6.4. Priorización de *hits ProP-PD* conteniendo SLiMs candidatos

Se utilizaron criterios para priorizar *hits* ProP-PD de Rb y p107 en el siguiente orden:

1. Detección de patrones de secuencia y anotación de variantes de mayor afinidad.
2. Evaluación de parámetros estructurales:
    a. Valor de accesibilidad relativa al solvente (RSA)
    b. Valor de predicción del desorden con IUPred
    c. Solapamiento con residuos Pfam
3. Localización celular
4. Estabilidad energética de matrices FoldX.

Se reportan a continuación los péptidos elegidos por orden de priorización en base a la variante del SLiM que se espera sea de mayor afinidad y el menor valor de FoldX que se estima que tendrá mayor afinidad. De poseer localización citosólica solo se seleccionaron, mediante inspección manual, aquellos que poseen roles similares a los de las proteínas *pocket*.

i- Péptidos hit de Rb con SLiM LxCxE: De los 308 *hits* de la proteína Rb, 44 contienen un SLiM LxCxE de los cuales sólo uno tiene una *regex* de categoría 1 y ocho tienen una *regex* de categoría 2A/2B (Tabla S21, Anexo). Luego de evaluar los parámetros mencionados, es posible priorizar 28 péptidos no solapados de los cuales 23 presentan advertencias de IUpred, Pfam, localización citosólica o se encuentran por encima del umbral de FoldX de cinco definido para este SLiM. 11 de estos *hits*, ya fueron testeados en el laboratorio incluyendo uno de categoría 1 y cinco de categoría 2A/2B (Tabla 6.1).

En la Tabla 6.4 se destacan cinco candidatos priorizables. El péptido **ZN180_HUMAN** es el único que posee un SLiM LxCxE que es de mayor afinidad por Rb que los SLiMs Lx[AST]xE. En las variantes de SLiMs LxCxE no se tuvo en cuenta esta variable por el bajo número de SLiMs con cisteínas rescatados. Otro péptido candidato a priorización es el de la proteína **TAF1_HUMAN**, que presenta un SLiM no reportado LxAxE. Al haber sido discutido en la sección 6.2 (Tabla 6.2) no se incluyó en los cinco ejemplos de *hits* candidatos. La búsqueda de términos GO reveló que la proteína citosólica **SUSD6_HUMAN** está asociada a la respuesta al daño en el ADN y la proteína **PCNT_HUMAN** está asociada la organización de los microtúbulos en mitosis y meiosis.

Tabla 6.4. Péptidos *hit* de Rb con SLiM LxCxE priorizados para ser candidatos de ensayos experimentales.

Uniprot ID	Secuencia[#]	Variante[†]	RSA	IUPred[#][#]	Pfam[#]*	Localización Celular**	FoldX***	Criterio aplicado
RBM33_HUMAN	EEQ**LYTDEVLD**IEINE	2A	0,84	0,72	16	NC	0,84	Priorizable (Pfam)
SUSD6_HUMAN	A**LPSYEEAV**YGSSGHC	2A	0,64	0,36	0	C	5,81	Priorizable (IUPred, Localización, FoldX)
PCNT_HUMAN	SVQK**LLAAEQTV**VRDL	3	0,52	0,38	0	C	2,09	Priorizable (IUPred, Localización)
NPHP4_HUMAN	FQFS**LGSEEHLD**APTE	3	0,65	0,60	3	NC	2,50	Priorizable
ZN180_HUMAN	T**LLCLEESM**EEQDEKP	3	0,90	0,69	0	N	2,85	Priorizable

[†]Variantes de regex definidas para el SLiM LxCxE en sección 3.2.1
[#]Subsecuencia que corresponde a variante de regex subrayada. Subsecuencia a la que corresponde el valor de FoldX resaltado en negrita.
[##] Una celda coloreada indica que existe una advertencia.
* Número de residuos solapados con dominios Pfam
**N= Núcleo; NC= Núcleo y Citoplasma
*** Valor de FoldX con la matriz FoldX 1GUX_8

ii- Péptidos hit de Rb en el ensayo ProP-PD con el SLiM E2F: De los 103 *hits* de la proteína Rb que contienen al SLiM E2F (Tabla S22, Anexo), se lograron priorizar 57 péptidos no solapados que cumplen con los criterios establecidos en este trabajo. De estos, 47 fueron identificados con advertencias de IUPred, Pfam, localización citosólica o FoldX, presentando un valor mayor a tres. De los 57 priorizables, 14 ya fueron testeados por el laboratorio. La Tabla 6.5 muestra el detalle de los cinco seleccionados por orden de priorización (lista completa en Tabla S22, Anexo). Según los términos GO, la proteína citosólica **AGGF1_HUMAN** está involucrada en la diferenciación celular y proliferación de células endoteliales.

Tabla 6.5. Péptidos *hit* de Rb con SLiM E2F priorizados para ser candidatos de ensayos experimentales.

Uniprot ID	Secuencia[#]	Variante[†]	RSA	IUPred	Pfam*	Localización Celular**	FoldX***	Criterio aplicado
HAX1_HUMAN	PALDDAFS**ILDLF**LGR	1	0,65	0,35	15	NC, SE	-1,64	Priorizable (IUpred y Pfam)
AGGF1_HUMAN	EVQTENHAPWS**ISDYF**	1	0,81	0,51	0	C	-0,63	Priorizable (Localización)
GON7_HUMAN	**VTELF**DPLVQGEVQHR	1	0,55	0,63	16	N	-0,31	Priorizable (Pfam)
SCAF8_HUMAN	**VFDYF**EGATSQRKGDN	1	0,86	0,68	0	N	1,16	Priorizable
VPS54_HUMAN	GMF**ISDAF**GEGELTPI	1	0,86	0,28	0	NC, SE	1,72	Priorizable (IUpred)

[#]Subsecuencia que corresponde a variante de regex subrayada. Subsecuencia a la que corresponde el valor de FoldX resaltado en negrita.
[†]Variantes de regex definida para el SLiM E2F en sección 3.2.2
* Número de residuos solapados con dominios Pfam
**N= Núcleo; NC= Núcleo y Citoplasma
***Valor de FoldX con la matriz FoldX 1N4M_5

iii- Péptidos hit de p107 en el ensayo ProP-PD con el SLiM LxCxE: Hay un total de 55 *hits* de la proteína p107 que contienen al SLiM LxCxE (Tabla S23, Anexo), de los cuales dos son variante 1 de *regex* y doce son variante 2A/2B. De ellos, se lograron priorizar 37 péptidos no solapados que cumplen con los criterios establecidos, de los cuales 13 fueron testeados experimentalmente incluyendo los dos con variante 1 de *regex* y cuatro con variante 2A/2B. 25 de los 37 péptidos presentan advertencias de IUPred, Pfam, localización celular o FoldX. En la Tabla 6.6 figuran los cinco seleccionados por orden de priorización. Las proteínas citosólicas **LDB3_HUMAN y PDLI1_HUMAN** tienen términos GO asociados a la diferenciación celular del músculo mientras que **MNAR1_HUMAN** está asociado a la regulación negativa de la diferenciación celular.

Tabla 6.6. Péptidos *hit* de p107 con SLiM LxCxE priorizados para ser candidatos de ensayos experimentales.

Uniprot ID	Secuencia[#]	Variante[†]	RSA	IUPred	Pfam*	Localización Celular**	FoldX***	Criterio aplicado
LDB3_HUMAN	QYNNPIG**LYSAETLR**E	2A	0,63	0,57	16	C	1,82	Priorizable (Pfam, Localización)
PDLI1_HUMAN	G**LYSSENIS**NFNNALE	2A	0,58	0,51	16	C	3,23	Priorizable (Pfam, Localización)
UTP25_HUMAN	S**LFSLETNF**LEEESGD	2A	0,57	0,63	0	N	3,63	Priorizable
MNAR1_HUMAN	KLTALD**LQTQESLN**PN	2B	0,67	0,45	16	C	2,35	Priorizable (Pfam, Localización)
ZN436_HUMAN	QWGD**LTAEEWVS**YPLQ	2B	0,71	0,4	0	NC	2,74	Priorizable

[#]Subsecuencia que corresponde a variante de regex subrayada. Subsecuencia a la que corresponde el valor de FoldX resaltado en negrita.
[†]Variantes de regex definidas para el SLiM LxCxE en sección 3.2.1
* Número de residuos solapados con dominios Pfam
**N= Núcleo; NC= Núcleo y Citoplasma
*** Valor de FoldX con la matriz FoldX 1GUX_8

iv- Péptidos hit de p107 en el ensayo ProP-PD con el SLiM E2F: Finalmente, de los 18 péptidos *hit* de p107 con el SLiM E2F detectado, diez péptidos no solapados cumplen con los criterios de priorización del presente trabajo (Tabla 6.7 y Tabla S24, Anexo) y seis de ellos fueron testeados en el laboratorio de trabajo. En la Tabla 6.7 figuran los cuatro mejores candidatos. Los términos GO asociados a la proteína citosólica **KIF15_HUMAN** indican funciones relacionadas con la migración cromosomal en mitosis mientras que los asociados a **NEBU_HUMAN** indican participación en la diferenciación celular y desarrollo de estructuras anatómicas.

Tabla 6.7. Péptido *hit* de p107 con SLiM E2F priorizado para ser candidato de ensayos experimentales.

Uniprot ID	Secuencia[#]	Variante[†]	RSA	IUPred[##]	Pfam[#*]	Localización Celular**	FoldX***	Criterio aplicado
KIF15_HUMAN	STQ**MQELF**SSERIDWT	1	0,57	0,46	0	C	0,56	Priorizable (Localización)
ZN436_HUMAN	QWGDLT**AEEWV**SYPLQ	3	0,71	0,4	0	NC	5,6	Priorizable (FoldX)
NEBU_HUMAN	KH**AMEVA**KKQSDVAYR	3	0,55	0,57	0	C	5,98	Priorizable (Localización, FoldX)
INKA1_HUMAN	LVLGDNCFADL**VHNWM**	4	0,54	0,32	16	NC	1,65	Priorizable (IUPred, Pfam)

[#]Subsecuencia que corresponde a variante de regex subrayada. Subsecuencia a la que corresponde el valor de FoldX resaltado en negrita.
[†]Variante de regex definida para el SLiM E2F en sección 3.2.2
[#] Una celda coloreada indica que existe una advertencia.
* Número de residuos solapados con dominios Pfam
**N= Núcleo; NC= Núcleo y Citoplasma
*** Valor de FoldX con la matriz FoldX 1N4M_5

Los péptidos que conforman estas listas, tienen un alto potencial para ser probados experimentalmente en el laboratorio, siendo todos ellos posibles interactores novedosos de las proteínas *pocket*. Como se mencionó previamente, aquellos que presentan advertencias deberán ser evaluados individualmente para luego considerarlos en ensayos experimentales.

Conclusiones de la aplicación de los criterios de priorización a hits ProP-PD provenientes de interactores conocidos reportados en IntAct o Proteómica: El análisis de los *hits* en conjunto con datos provenientes de ensayos de gran escala (ver Sección 6.2) permitió priorizar **un** péptido con un SLiM LxAxE y **tres** péptidos con un SLiM E2F canónico. Además, se identificaron **dos péptidos** solapados provenientes de la proteína PRUN2_HUMAN que presentan un SLiM *E2F-like* (Tabla 6.2) que actualmente están siendo ensayados en el laboratorio. Por último, se identificó en la proteína **PLAK2_HUMAN** un péptido a priorizar sin SLiM detectado pero con características de secuencias similares al péptido KNL1_HUMAN previamente ensayado en el laboratorio.

De los **seis** péptidos a priorizar cinco poseen localización celular citoplasmática. Sin embargo estos péptidos podrían estar involucrados en funciones relacionadas con las proteínas *pocket*, como ocurre en el caso del *hit* **SP** testeado en el laboratorio de la kinesina KIF24_HUMAN (Tabla 6.1) involucrada en procesos mitóticos, ya que durante la mitosis las proteínas citosólicas podrían interactuar con Rb luego de la disgregación de la membrana nuclear. En el futuro, el uso de términos GO será incorporado en el análisis para facilitar la identificación de proteínas con roles similares a las proteínas *pocket*.

Conclusiones de la aplicación de los criterios de priorización a hits ProP-PD con SLiMs identificados por MEME. Se incluyen como péptidos de interés a futuro los péptidos con SLiMs

identificados utilizando la herramienta MEME: **dos péptidos** en los que se identificó un SLiM *E2F-like* con una variante no reportada previamente y no incluida en la *regex* usada en este trabajo y **un péptido** en el que se detecta tanto el SLiM E2F "reverso" como un SLiM *E2F-Like* similar al presente en KNL1_HUMAN (Tabla 6.3). En ambos casos, los SLiMs identificados podrían representar modos de interacción novedosos que es necesario evaluar en futuros ensayos experimentales.

Concluímos que el ensayo ProP-PD permite mapear SLiMs no reportados en interactores conocidos de las proteínas *pocket*, así como identificar posibles variantes novedosas de SLiMs para esta familia de proteínas. Por este motivo, será de interés futuro del laboratorio, evaluar con estos criterios aquellos péptidos que fueron *hit,* pero que no presentan las expresiones regulares definidas, teniendo en cuenta aquellos cuyos parámetros estructurales y/o de estabilidad de secuencia pasen los filtros y umbrales establecidos en el presente trabajo.

Conclusiones de la aplicación de los criterios de priorización a hits ProP-PD con SLiMs conocidos:
Se destacó al principio del capítulo la importancia de realizar una evaluación global que incluya más de un parámetro para establecer un criterio de priorización para seleccionar péptidos candidatos a ensayos experimentales. Se consideraron además péptidos de localización citosólica que cumplen con los parámetros definidos y que se encuentran involucrados en funciones asociadas a las de la familia *pocket*. Al emplear estos criterios, se lograron priorizar un total de **85 péptidos** que fueron *hit* de Rb: **28 con SLiM LxCxE** (Tabla Suplementaria S21, Anexo) y **57 con el SLiM E2F** (Tabla Suplementaria S22, Anexo). Para la proteína *pocket* p107, se priorizaron un total de **47 péptidos** *hit*: **37 con SLiM LxCxE** (Tabla Suplementaria S23, Anexo) y **diez con SLiM E2F** (Tabla Suplementaria S24, Anexo). Del total de péptidos priorizados para ambas proteínas y SLiMs, se seleccionaron 19 *hits* a ser evaluados en ensayos experimentales.

6.5. Conclusión General

Si bien las técnicas de gran escala arrojan un alto número de interactores, los ensayos *in vitro* no necesariamente identifican interactores fisiológicamente relevantes dado que variables como la localización celular, co-expresión en un mismo tipo celular o estado funcional de la célula o la accesibilidad de los SLiMs en cada proteína pueden condicionar que ocurra o no una interacción. En este trabajo se utilizó una lista de péptidos provenientes de una biblioteca construida con el desorderoma humano, que se unieron a los dominios *pocket* de Rb y p107 en el ensayo ProP-PD, aunque en un contexto no fisiológico [40].

Dada la baja complejidad y corta secuencia de los SLiMs, resulta probable encontrarlos en diferentes IDRs, formando parte de proteínas no relacionadas funcionalmente con la familia *pocket*, o bien en un contexto proteico poco accesible para interactuar con otros blancos proteicos. En estos casos, los SLiMs no son necesariamente funcionales. Es por ello que se abordó el análisis de péptidos *hit* desde un punto de vista estructural, de secuencia y estabilidad energética utilizando herramientas bioinformáticas que nos permitan en conjunto, diseñar una estrategia de filtrado y priorización de candidatos que reúnan condiciones favorables de interacción. Esta priorización genera un mayor nivel de confianza en los candidatos a ser testeados experimentalmente y a la vez logrará optimizar los recursos disponibles en el laboratorio. En total, se seleccionaron 28 péptidos *hits* a ser validados experimentalmente tomando en cuenta los diferentes criterios utilizados.

Finalmente, además de la priorización de péptidos realizada para evaluar en futuros ensayos experimentales, las herramientas desarrolladas para el procesamiento y priorización de los datos del cribado de la biblioteca HD2 con Rb y p107 *wild-type* será utilizado para evaluar los resultados de otros 14 cribados realizados disponibles en el laboratorio incluyendo un nuevo cribado de alta calidad realizado con p130, cribados realizados utilizando mutantes del bolsillo LxCxE de Rb y p107, y cribados que utilizan únicamente las sub-librerías de interés: nuclear (N) y núcleo-citoplasmática (NC) que ayudan a priorizar proteínas funcionalmente relevantes evitando la competición con blancos espúreos.

Los resultados presentados, junto con la información disponible sobre SLiMs de interacción con proteínas *pocket*, contribuyen a ampliar el conocimiento sobre las bases moleculares de las interacciones que median procesos regulatorios cruciales, como la regulación de la proliferación y diferenciación celular y la progresión del cáncer.

Capítulo 7: Métodos

Los scripts utilizados para el análisis de datos fueron realizados en Python3 y R v3.4.2. Los gráficos se realizaron en R utilizando el paquete ggplot2 v3.4. Los mismos se encuentran en el siguiente link: https://github.com/chemeslab/Lorenze_Carla.

7.1. Descripción de la construcción de la biblioteca HD2

El primer paso del ensayo ProP-PD fue la construcción de la biblioteca HD2 (*Human Desorderome 2*) [40] (Figura 7.1, izquierda). Esto fue realizado por los grupos colaboradores del Dr. Norman Davey (Inglaterra) y la Dra. Ylva Ivarsson (Suecia). A partir de proteínas que pertenecen al proteoma humano, en primer lugar se redujo el espacio de búsqueda según anotaciones Uniprot a proteínas intracelulares con regiones desordenadas y *loops* de regiones estructuradas, descartando proteínas relacionadas con los términos "secreción", a menos que también se encuentren relacionadas con los términos "citoplasma" y/o "núcleo", "regiones transmembrana" y "regiones extracelulares" [65].

Luego, utilizaron tres fuentes de datos analizados de manera jerárquica para definir regiones intrínsecamente desordenadas (IDRs) y *loops* de regiones estructuradas:

1. accesibilidad al solvente (SA) de estructuras resueltas de proteínas,
2. accesibilidad al solvente de proteínas homólogas con estructura resuelta, y
3. predicción del desorden con IUPred

Para aquellas proteínas con estructura resuelta o bien homólogas a éstas (puntos 1 y 2), se calculó la accesibilidad al solvente y se clasificaron en accesibles con un valor mayor al umbral de 33%. Para aquellas proteínas sin información experimental, se utilizó el predictor de desorden IUPred tomando regiones desordenadas por encima del umbral de 0.4.

Por último utilizaron un criterio de suavizado (*smoothing*) que permitió definir el espacio de búsqueda:

- descartaron regiones de menos de 4 residuos que no son consistentes con la categoría de las regiones flanqueantes,
- conservaron regiones ordenadas de menos de 25 residuos en un contexto desordenado, y
- conservaron péptidos de 16 residuos con al menos 8 residuos clasificados como accesibles

Se definieron péptidos de 16 residuos con 12 residuos de solapamiento. Aquellos péptidos de loops citoplasmáticos de 8 residuos o más, se consideraron desordenados y fueron incluidos en el espacio de búsqueda de ProP-PD. Una vez definidos los péptidos, todas las cisteínas (C) fueron mutadas a alaninas (A) para evitar oxidación de cisteínas y formación de puentes disulfuro en la superficie de los

fagos.

El archivo de la biblioteca HD2 en formato de tabla fue proporcionado por los laboratorios del Dr. Norman Davey y la Dra. Ylva Ivarsson; y fue el punto de partida del presente trabajo.

7.1.1. Organización de la biblioteca

La biblioteca HD2 está organizada en pooles de péptidos según la localización subcelular de anotaciones en Uniprot de la proteína de la que provienen y términos GO [65,66]. La biblioteca HD2 se dividió en distintas sub-bibliotecas con el fin de reducir el número de interacciones competitivas. El tamaño de cada archivo se corresponde con la cantidad de espacios libres del chip comercial utilizado por el grupo de trabajo para realizar el ensayo ProP-PD. Se definieron entonces, sub-bibliotecas de diferentes tamaños: (1) citoplasma, sin incluir compartimento nuclear- 3 archivos de 92918 péptidos, (2) núcleo sin citoplasma- 3 archivos de 92918 péptidos, (3) sistema endomembrana- 2 archivos de 92918 péptidos, (4) núcleo y citoplasma- cuatro archivos de 92918 péptidos, (5) extracelular- 1 archivo de 92918 péptidos, (6) otros- 1 archivo de 92918 péptidos. Los espacios vacíos de los chips comerciales fueron llenados con péptidos redundantes de cada sub-biblioteca.

7.1.2. Transformación de fagos, incubación y selección

El equipo de trabajo obtuvo los oligonucleótidos codificantes de los péptidos de la biblioteca HD2 que fueron desplegados en un sistema de fagos M13 [40]. Según protocolo previamente descrito [67,68] realizaron la transformación de fagos, incubación y selección de péptidos expuestos en la superficie de los fagos (Figura 7.1, arriba). Luego de cada selección, se secuencian (Illumina) las regiones codificantes de los pools de fagos enriquecidos para evaluar los péptidos de interacción con los dominios de interés (Figura 7.1).

Finalmente se obtuvo una lista de péptidos *hit* con información sobre el número de réplicas en la cual se recupera el péptido, número de cuentas totales, es decir el número total de veces que el péptido es detectado por NGS en el total de réplicas y número de péptidos solapantes obtenidos dentro del conjunto de *hits* [40].

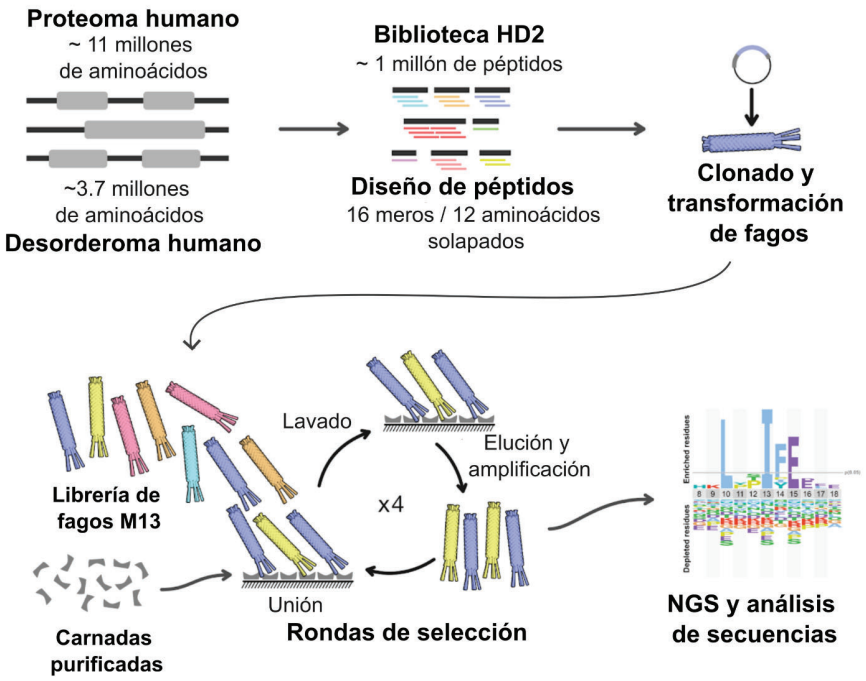


Figura 7.1. Esquema de la técnica de ProP-PD. Arriba: biblioteca de fagos de regiones desordenadas del proteoma humano expresados como péptidos de 16 aminoácidos, de los cuales 12 residuos se encuentran solapados con péptidos de una región contigua, logrando así cubrir con más de una copia los segmentos desordenados a analizar. Los péptidos se retro-transcribieron y añadieron adaptadores para construir una biblioteca de oligonucleótidos de fagos. Abajo: la biblioteca de fagos fue sometida a un proceso de selección iterativo (rondas de selección, recuadro punteado), donde se la incubó con los dominios *pocket* de Rb, p107 y p130. Los péptidos enriquecidos durante la selección fueron secuenciados y analizados por NGS. Paneles adaptados de Benz et al. 2022 [40].

7.1.3. Métricas de calidad establecidas

Para evaluar la calidad del ensayo ProP-PD, los investigadores definieron métricas de calidad que utilizaron para evaluar a los péptidos enriquecidos en las rondas de selección. Utilizaron un conjunto de datos comparativo de SLiMs conteniendo 466 instancias de SLiMs previamente validados en ELM [36] como interactores de las 40 proteínas carnada que se utilizaron en el ensayo ProP-PD. De estas instancias, 337 se encontraban presentes en la biblioteca HD2 diseñada e interaccionan con al menos una de las 35 carnadas no utilizadas como control [40]. Aquellos péptidos solapados con instancias validadas del conjunto de datos comparativo de SLiMs, se compararon con los péptidos seleccionados en las diferentes rondas del ensayo para cada carnada que no pertenecen al set de datos validado. Las cuatro métricas utilizadas para comparar péptidos fueron:

1. Péptidos replicados: número de réplicas en las que se encontró al péptido
2. Péptidos solapados, el número de péptidos distintos que solapan con otros péptidos en las réplicas
3. Coincidencia de SLiMs entre los péptidos seleccionados en las réplicas y las instancias definidas en ELM [36]

4. Cuentas de péptidos normalizados: el promedio de número de cuentas para cada péptido que surge del análisis por NGS de cada réplica.

7.2. Organización de bases de datos

Se unificaron los 14 archivos de las sub-bibliotecas de la base de datos HD2 y se eliminaron las filas conteniendo péptidos redundantes en un archivo final de 941618 péptidos únicos en base a la secuencia con la mutación por alanina (A). El archivo de texto plano final en formato '.csv' de la biblioteca HD2 contiene información de los péptidos únicos incluidos en la librería indicando: la secuencia con cisteína y la secuencia mutada por alanina, el número de acceso, UniRef (referencia de clusters de Uniprot), identificador Uniprot, nombre completo de la proteína, total de residuos de la proteína, cantidad de residuos desordenados en la proteína completa, porcentaje de residuos desordenados en la proteína completa, posiciones de inicio y fin del péptido, si la secuencia contiene una mutación de C→A y localización celular.

Se creó un identificador único de cada péptido conformado por tres valores: número de acceso, posición de inicio y secuencia no mutada.

Para cada péptido se incluyó además si era un "verdadero positivo" para los SLiMs LxCxE y E2F según los datos recolectados de la base de datos ELM [36] 2021. Los datos de ELM se complementaron con "verdaderos positivos" recolectados de la literatura: E2F3, E2F4, E2F5 y LIN52.

También se incluyó para cada péptido información proveniente de la base de datos de interactores IntAct [42], indicando si el péptido pertenece a una proteína reportada como interactor directo/indirecto de Rb, p107 o p130. Por último, para cada péptido se indicó si pertenecían o no a una proteína reportada en un ensayo de proteómica con Rb [48].

7.2.1. Lista de péptidos hit para cada proteína *pocket*

A partir del ensayo ProP-PD realizado por el grupo de la Dra. Ivarsson, se obtuvo una lista de péptidos que fueron *hits* para cada uno de los dominios *pocket* utilizados como carnada. Estos archivos contienen información sobre el nombre de la proteína, la secuencia wild-type del péptido que fue *hit*, secuencia con la mutación C→A, cantidad de réplicas realizadas en el ensayo para cada péptido, solapamiento del péptido (12 residuos o mas) con otro péptido *hit* de la lista, número de cuentas luego del análisis de secuenciación para cada réplica y sumatoria total de cuentas para cada péptido, el número de acceso de la proteína a la que pertenece el péptido, nombre del gen y posiciones de inicio y fin del péptido en la secuencia proteica [54].

7.3. Evaluación de calidad del ensayo ProP-PD

Con el fin de evaluar la recuperación de interactores validados en el ensayo ProP-PD, se realizó una búsqueda de SLiMs previamente reportados e interactores conocidos en las bases de datos: ELM (Eukaryotic Linear Motif database) [36], IntAct (Molecular Interaction database) [42] y en ensayos de proteómica reportados previamente para Rb [48].

7.3.1. Búsqueda en bases de datos y datos de proteómica: Identificadores de las proteínas *pocket*

Los dominios *pocket* de las tres proteínas de la familia, fueron utilizados como carnada en el ensayo ProP-PD. En la Tabla 7.1 se reportan los identificadores correspondientes utilizados a lo largo de todo el trabajo.

Tabla 7.1. Identificadores de las proteínas pocket.

Nombre de la proteína	Retinoblastoma-associated protein pRb Rb	Retinoblastoma- like protein 1 pRb1 p107	Retinoblastoma- like protein 2 pRb2 p130
Identificador Uniprot	RB_HUMAN	RBL1_HUMAN	RBL2_HUMAN
Nombre del gen	RB1	RBL1	RBL2
Número de acceso	P06400	P28749	Q08999

7.3.2. Recolección de Verdaderos Positivos: ELM, Base de datos de Motivos Lineales (Eukaryotic Linear Motif database)

La base de datos ELM reúne anotaciones de SLiMs eucariotas validados experimentalmente que fueron curados manualmente a partir de literatura. Las instancias presentes en ELM se clasifican según el tipo de SLiM, su función y clase ELM, esta última descrita por una expresión regular. Las instancias de cada clase ELM se clasifican en: (1) verdadera positiva (TP, ***T**rue **P**ositive*) si la instancia posee evidencia experimental que muestra que es funcional; (2) falso positivo (FP, ***F**alse **P**ositive*), si existe una instancia con evidencia experimental que sugiere una función pero luego de revisión por parte de los curadores, se considera una instancia no funcional; (3) verdadero negativo (TN, ***T**rue **N**egative*), si la evidencia experimental demuestra que no es funcional y (4) desconocido (U, ***U**nknown*) , si no existe información que determine si es funcional o no [36].

Se recolectaron todas aquellas instancias correspondientes a los SLiMs de interacción con las proteínas *pocket* identificadas como “LIG_Rb_LxCxE_1” y “LIG_Rb_pABgroove_1” que correspondieran a *Homo sapiens*. Al momento de la búsqueda (noviembre 2021) se encontraron quince instancias TP en total para ambos SLiMs. Estos datos se complementaron con cuatro instancias recolectadas de literatura, alcanzando un total de 19 instancias [34]: la proteína LIN-52 (LIN52_HUMAN, Q52LA3) con un SLiM LxSxE [22], y tres proteínas con un SLiM E2F: el factor de transcripción E2F3 (E2F3_HUMAN, O00716) [69], el factor de transcripción E2F4

(E2F4_HUMAN, Q16254) [69] y el factor de transcripción E2F5 (E2F5_HUMAN, Q15329). Hasta la fecha (agosto 2024), con excepción de las instancias recolectadas de la literatura, no se reportaron en ELM nuevos SLiMs de interacción correspondientes a humanos para los SLiMs E2F y LxCxE (Tabla S1, Anexo) [38].

La información recolectada incluye el identificador de ELM de cada instancia, el tipo de interacción bajo el que fueron anotadas (LIG: ligando, en este caso), nombre de la proteína, número de acceso, posiciones de inicio y fin del péptido en la proteína, referencias de la publicación donde se identificó la interacción, el método mediante el cual se identificó, si la interacción fue clasificada como TP, FP, TN o U, el identificador PDB, el organismo del que proviene la instancia anotada y una secuencia de 20 aminoácidos conteniendo el SLiM de interacción.

Se identificó y anotó la presencia de verdaderos positivos (reportados con SLiMs 'LxCxE', 'E2F' o 'LxSxE') en la biblioteca HD2 y en los *hits* del ensayo ProP-PD para las proteínas *pocket*, buscando coincidencias con la tabla recolectada utilizando dos criterios: coincidencia en los números de acceso y solapamiento del 50% de la secuencia aminoacídica (8 residuos), teniendo en cuenta las posiciones de inicio y fin del péptido en cada caso. Este umbral fue definido tomando como referencia el solapamiento de proteínas conocidas conteniendo SLiMs de interacción validados [36].

Los datos recolectados nos permitieron estimar el *recall* global del ensayo y el *recall* para cada proteína *pocket*. El *recall* es el porcentaje de instancias conteniendo SLiMs conocidos recuperados en relación con la cantidad de instancias presentes en la biblioteca HD2 y se calculó de la siguiente manera:

$$Recall = \frac{interactores\ recuperados}{interactores\ presentes\ en\ HD2} \times 100$$

7.3.3. Base de datos IntAct

IntAct es una base de datos pública de interacciones moleculares, que reúne datos seleccionados de literatura o directamente enviados por los usuarios [42]. Para realizar la búsqueda, se utilizaron los números de acceso de las proteínas *pocket* y se aplicaron los siguientes filtros:

- **Especies interactoras**: se buscaron interacciones exclusivas de *Homo sapiens* con intra-interacciones (provenientes de proteínas de Homo sapiens con proteínas de la misma especie) dado que, tanto los dominios *pocket* utilizados para realizar el ensayo ProP-PD como el objeto de estudio del presente trabajo, son proteínas humanas con interacciones en su propio contexto celular;
- **Tipo de interactor**: proteico, dado que es de nuestro interés identificar interacciones específicas mediadas por SLiMs con la familia de proteínas *pocket*;
- **Tipo de interacción**:

- *Asociación física o indirecta*: cuando los interactores fueron identificados dentro de un complejo, es decir, dentro de un grupo de moléculas en contacto con la proteína de interés
- *Interacciones directas*: incluye fosforilación, metilación, acetilación y desacetilación, además de interacciones entre proteínas.

Se obtuvieron 24 interactores de Rb de tipo de interacción directa y 78 indirectos (Tabla S3, Anexo), para p107 se encontraron tres interactores directos y 35 indirectos (Tabla S4, Anexo) y en el caso de p130 se encontraron tres interactores directos y 13 indirectos. Se identificó y anotó la presencia de verdaderos positivos en la biblioteca HD2 y en los hits del ensayo ProP-PD, buscando coincidencias con la tabla recolectada utilizando los números de acceso. Se anotó además cuáles de ellos son interactores directos o indirectos de Rb, p107 y/o p130.

Se estimó el *recall* de interactores recolectados en la base de datos IntAct teniendo en cuenta el número de interactores recuperados luego del cribado y de presentes en la biblioteca HD2, de la siguiente manera.

$$Recall = \frac{interactores\ recuperados}{interactores\ presentes\ en\ HD2} \times 100$$

Comparación con técnicas de gran escala

Se contrastaron los resultados obtenidos de ProP-PD con 438 interactores reportados en otra técnica de gran escala, en este caso proteómica, donde se evaluó la interacción con Rb monofosforilada en distintos sitios [48].

Se identificaron y anotaron aquellos presentes en la biblioteca HD2 y entre los péptidos que fueron hits en ProP-PD buscando coincidencia en el número de acceso.

7.4. Detección de patrones de secuencia en péptidos hits

Los SLiMs pueden representarse mediante expresiones regulares que surgen de los patrones de secuencia que se observan en el alineamiento de instancias conocidas del SLiM y de información inferida a partir de la estructura de los complejos SLiM-dominio.

Con el objetivo de identificar los SLiMs conocidos E2F y LxCxE presentes en péptidos *hit* de Rb y p107 en el ensayo ProP-PD, se desarrolló un programa de identificación de las variantes consideradas de cada SLiM que permitió separarlos en tres grupos: 1) péptidos con SLiM LxCxE, 2) péptidos con SLiM E2F y 3) péptidos sin SLiM detectado.

7.4.1. SLiM LxCxE

El SLiM LxCxE se encuentra definido en la base de datos ELM [36] de la siguiente manera:

```
[DEST].{0,4}[LI].C.E.{1,4}[FLMIVAWPHY].{0,8}[DEST]
```

El centro o *core* del SLiM LxCxE se encuentra definido por cinco residuos, donde la primera, tercera y quinta posición son fijas (indicadas en azul) y las posiciones dos y cuatro son variables (indicadas por un punto).

- En la **primera posición fija** se admite un residuo I o L, presentes en instancias reportadas, como KDM5A_HUMAN (LxCxE) o HDAC_HUMAN (IxCxE) [36].
- En la **segunda posición fija** se admite la cisteína C y serina S, reportada como variante del SLiM en la proteína LIN52_HUMAN (LxSxE). Se incluyó también al residuo A, cuya afinidad de interacción disminuye 62 veces la afinidad de interacción , mientras que la afinidad del residuo S disminuye 220 veces (Tabla S2 Anexo) en el péptido con el SLiM LxCxE de de alta afinidad de la proteína E7 del papilomavirus. Por último en esta posición se consideró una treonina T por su similitud estructural con el residuo S.
- En la **tercera posición fija** se consideró únicamente al glutámico E.

Si bien las posiciones variables dos y cuatro admiten cualquier residuo, se observó que los residuos aromáticos fenilalanina, triptófano y tirosina se posicionan hacia afuera del dominio, aumentando la afinidad de interacción [34,50].

Dado que para la construcción de la biblioteca HD2 se realizó la mutación C→A, esto afecta al SLiM en la **segunda posición fija** disminuyendo la afinidad de unión del péptido por el dominio *pocket* (Tabla S2, Anexo). Por esto, se buscó identificar péptidos que pudiesen tener alta afinidad de unión a pesar de no poseer la C central, restringiendo la expresión regular a residuos preferidos en las posiciones variables del **core** y flanqueantes al **core** del SLiM, siendo éstos:

- ácidos: aspártico (D) y glutámico (E) en la posición -1 hacia el extremo N-terminal
- aromáticos: fenilalanina (F), triptofano (W), tirosina (Y) en las posiciones variables del centro del SLiM
- hidrofóbicos: triptofano (W), fenilalanina (F), isoleucina (I), leucina (L), tirosina (Y), valina (V), metionina (M) en las posiciones +2 o +3 hacia el extremo C-terminal.

En las posiciones fijas del ***core*** del SLiM:

- Se admitieron los residuos I y L en la primera posición fija
- Cisteína y serina (S) en la segunda posición fija, por ser variantes del SLiM, A por presentar mayor afinidad de unión que S (Tabla S2, Anexo)) y treonina (T) por su similitud estructural con S.

El segmento acídico posicionado en el extremo C-terminal no fue considerado para el análisis dado que los péptidos de 16 residuos no alcanzan la longitud mínima necesaria que lo contiene.

Para detectar estas características se utilizaron variantes de expresiones regulares del SLiM LxCxE y se agruparon en cinco categorías por orden de prioridad, de acuerdo a las características que mayor afinidad le confieren a la interacción entre el péptido y la proteína *pocket,* donde 1 es la

categoría con mayor afinidad de unión y 5 con menor afinidad (Tabla 7.2).

Tabla 7.2. Definición de variantes del SLiM LxCxE.

Categoría#	Variante LxCxE	Expresión regular
1	Res. ácido en -1, posición/es variable/s, res. hidrofóbico +2/+3	**[DE]**[IL]**[YFH]**[CAST].E.**{1,2}[WFILYVM]** **[DE]**[IL].[CAST]**[YFH]**E.**{1,2}[WFILYVM]**
2	**A.** Posición/es variable/s, res. hidrofóbico +2/+3	[IL]**[YFH]**[CAST].E.**{1,2}[WFILYVM]** [IL].[CAST]**[YFH]**E.**{1,2}[WFILYVM]**
	B. Res. ácido en -1, res. hidrofóbico +2/+3	**[DE]**[IL].[CAST].E.**{1,2}[WFILYVM]**
	C. Res. ácido en -1, posición/es variable/s	**[DE]**[IL]**[YFH]**[CAST].E **[DE]**[IL].[CAST]**[YFH]**E
3	Res. hidrofóbico +2/+3	[IL].[CAST].E.**{1,2}[WFILYVM]**
4	**A.** Res. ácido en -1	**[DE]**[IL].[CAST].E
	B. Posición/es variable/s	[IL]**[YFH]**[CAST].E [IL].[CAST]**[YFH]**E
5	Res. centrales	[IL].[CAST].E

#Las categorías se listan en orden decreciente de afinidad.

7.4.2. SLiM E2F

La expresión regular del SLiM E2F se encuentra definida en ELM [36] como:

..[LIMVA].[DE][LMF][FYM][IL]{0,1}([DE]|(S)).

- La **primera posición fija [LIMVA]**: se la mantuvo como la describe ELM.
- La siguiente **posicion variable**: se mantuvo como la describe ELM.
- La **segunda posición fija [DE]**: se incluyó también la glutamina (Q) y asparagina (N), dado que son residuos polares que pueden establecer puentes de hidrógeno con residuos del dominio *pocket*, así como lo hacen como D y E,
- En la **cuarta y quinta posición fijas:** se incluyeron todos los residuos aromáticos e hidrofóbicos alifáticos.
- La última posición hidrofóbica [IL]{0,1}, así como la posición ácida [DE]|(S) no se incluyeron en la definición.

Esto permitió hacer la expresión regular más abarcativa, dada la falta de información disponible para este SLiM.

Se utilizaron variantes de expresiones regulares del SLiM E2F y se agruparon en cuatro categorías por orden de prioridad, de acuerdo a las características que mayor afinidad le confieren a la interacción entre el péptido y la proteína *pocket,* donde 1 es la categoría con mayor afinidad de unión y 5 con menor afinidad (Tabla 7.3).

Tabla 7.3. Definición de variantes del SLiM E2F.

Categoría[#]	Variante E2F	Expresión regular
1	Res. ácido en tercera posición, res. F/Y en quinta posición	`[IVLMA].[DE][IVLFMYAW][FY]`
2	Res. F/Y en quinta posición	`[IVLMA].[NQDE][IVLFMYAW][FY]`
3	Res. ácido en tercera posición	`[IVLMA].[DE][IVLFMYAW][IVLFMYAW]`
4	Res. centrales	`[IVLMA].[NQDE][IVLFMYAW][IVLFMYAW]`

[#]Las cuatro categorías se listan en orden decreciente de afinidad.

7.4.3. Enriquecimientos de SLiMs

Con el fin de evaluar la sobrerrepresentación de SLiMs presentes en los péptidos *hit* de Rb y p107 de manera no sesgada, se utilizó el servidor web MEME [52] en el total de péptidos *hit* de 16 residuos de las selecciones ProP-PD de ambas proteínas. El algoritmo de MEME emplea una técnica donde maximiza expectativas de ocurrencias en un conjunto de secuencias, asociando de a dos componentes por vez y clasificando probabilísticamente las apariciones del SLiM identificado. Esto se realiza de manera repetitiva para encontrar SLiMs [70]. Los datos de entrada son un conjunto de secuencias sin gaps de más de 8 residuos y busca un patrón compartido entre todas ellas.

Además de evaluar la sobrerrepresentación de SLiMs en el total de péptidos *hit* de cada proteína, se analizaron por separado los grupos de péptidos:

1. **Con SLiM LxCxE:** péptidos cuya secuencia contiene al menos una variante del SLiM LxCxE
2. **Con SLiM E2F:** péptidos cuya secuencia contiene al menos una variante del SLiM E2F
3. **Sin SLiM:** péptidos cuya secuencia no contiene variantes de los SLiMs LxCxE y/o E2F

Para realizar esta búsqueda se modificó el parámetro del largo del SLiM a identificar entre 5 y 16 posiciones, teniendo en cuenta que tanto el SLiM LxCxE como el E2F tienen un largo mínimo de 5 residuos, y los péptidos *hit* son de 16 residuos. El resto de los parámetros utilizados son por defecto los que define el sitio web. Para cada grupo se seleccionaron los logos de menor E-valor aceptándose como máximo un E-valor de 0.005 para que el logo sea significativo.

7.5. Filtrado y priorización de péptidos *hit* según parámetros estructurales

7.5.1. Accesibilidad relativa al solvente

El área de la superficie accesible al solvente (SASA, ***S**olvent-**A**ccessible **S**urface **A**rea*) es un valor que estima la exposición y, por lo tanto, accesibilidad de un residuo para interaccionar con el solvente. El valor se obtiene considerando la superficie proteica a partir de una estructura, y la distancia entre una molécula de agua de un radio de 1.5Å y un residuo "X" en el contexto proteico

(algoritmo de Shrake–Rupley) [57].

La accesibilidad relativa al solvente (RSA) es el SASA de cada residuo relativizado al SASA del residuo "X" en contexto de un pentapéptido conformado por glicinas (GG-X-GG). Este valor relativo de accesibilidad al solvente va desde 0 cuando el residuo es inaccesible hasta 1 cuando el residuo se encuentra totalmente expuesto y accesible para interactuar [54] (Figura 7.2).

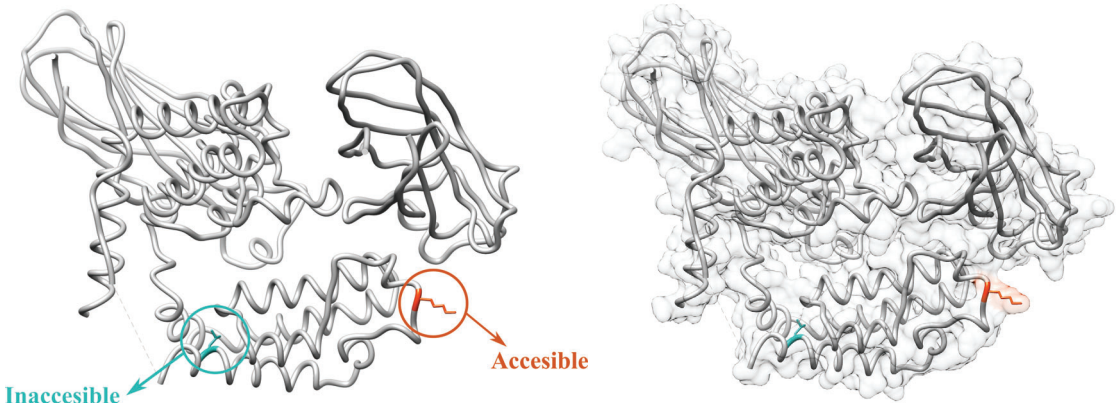


Figura 7.2- Comparación de residuos accesibles e inaccesibles de TFIID (UniprotID: P21675). Estructura del complejo de doble bromodominio CIA/ASF1 de histona chaperona que vincula modificaciones de histona y desalojo sitio-específico (PDB: 3AAD). Se señala en color anaranjado una lisina (K), accesible y expuesta para interactuar ubicada en un loop desordenado, y en verde un residuo leucina (L) inaccesible, orientado hacia el seno de la región estructurada (izquierda). Representación de la superficie del complejo, donde se visualiza el residuo lisina accesible, mientras que la leucina se localiza en el seno de la estructura alfa hélice, quedando inaccesible para interactuar con targets proteicos (derecha).

A partir de estructuras modeladas en Alphafold2 [54], obtenidas de la base de datos AlphaFold Protein Structure Database en Octubre 2022, se calculó el valor de SASA utilizando el programa DSSP v3.0.0 [71] y luego se obtuvo el valor de RSA para cada proteína. Por último, se calculó un valor de RSA promedio para cada uno de los péptidos *hit* de 16 residuos utilizando la posición de inicio y fin correspondientes.

Si bien IUPred es un predictor que correlaciona fuertemente con evidencia experimental y es ampliamente utilizado en la literatura para la predicción de desorden [55], el uso de herramientas adicionales como el RSA obtenido a partir de modelos AlphaFold o estructuras cristalinas permitiría una mayor confianza en la predicción. Por ejemplo, los péptidos hits provenientes de las proteínas VWA3B_HUMAN y DACT1_HUMAN poseen un mismo valor de IUPred (0,3), sin embargo, el valor de RSA permite distinguir que el péptido de DACT1_HUMAN está accesible mientras que el péptido de VWA3B_HUMAN no está accesible (Figura 7.3) mejorando la priorización. Por ejemplo, se identificaron péptidos con mismo valor de IUPred indicando mismo grado de desorden y un valor de RSA distinto obteniendo un criterio para la priorización (Figura 7.3).

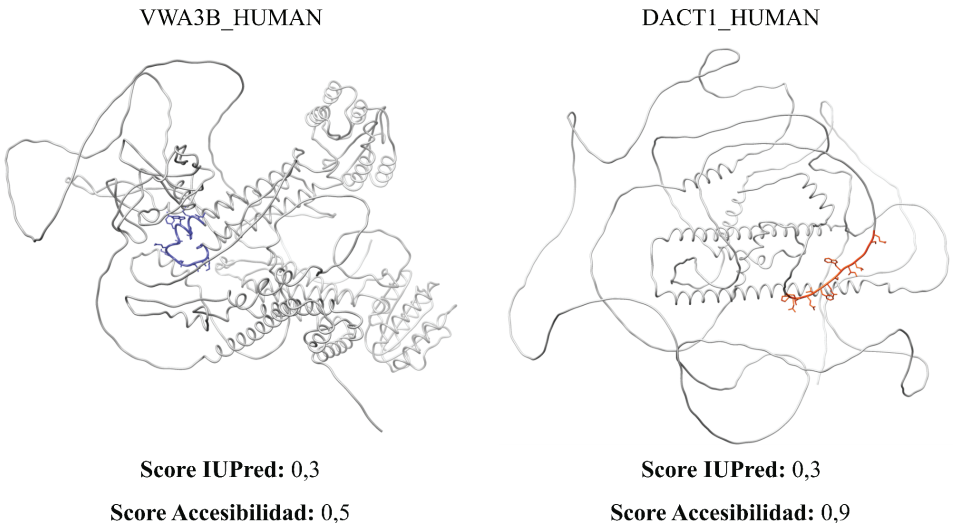


Figura 7.3. El análisis estructural de péptidos hits debe incluir diferentes parámetros. Estructuras de modelos Alphafold de las proteínas VWA3B_HUMAN (UniprotID:Q502W6) (izquierda) y DACT1_HUMAN (UniprotID: Q9NYF0) (derecha). Se señalan en azul y rojo, respectivamente, los péptidos de 16 residuos que fueron hits en ProP-PD utilizando el dominio de Rb como carnada. Si bien los valores de IUPred son iguales, sugiriendo un grado de desorden similar, se observa que el péptido de VWA3B_HUMAN se encuentra en el seno de una región estructurada con un valor de accesibilidad bajo, mientras que el de DACT1_HUMAN se ubica en una región accesible y expuesta para interactuar, con un score de accesibilidad alto.

Estas observaciones sugieren que el análisis de péptidos mediante RSA proporciona una estimación más precisa como parámetro estructural en comparación con IUPred. Si bien al momento de diseñar la biblioteca se utilizaron las estructuras de proteínas o proteínas homólogas para seleccionar péptidos accesibles, no hay estructuras reportadas para muchas de las proteínas incluidas en la biblioteca. El uso de modelos AlphaFold2, que no estaban disponibles al momento de diseñar la biblioteca, representa una mejora al momento de la selección de péptidos *hits*.

7.5.2. Predicción del desorden (IUPred)

Se utilizó el predictor de desorden IUPred2A (IUPred en adelante) para realizar un primer análisis estructural. IUPred predice regiones intrínsecamente desordenadas (IDRs) en base a la energía de un residuo en el contexto de aminoácidos en el que se encuentra. El score de IUPred es obtenido a partir de un valor promedio de contribuciones energéticas de aminoácidos vecinos en una ventana de "n" residuos, donde el resultado es transformado a un valor de IUPred entre 0 y 1 [55]. Los valores entre 0 y 0.5 indican que el residuo se encuentra en una región ordenada y entre 0.5 y 1 en regiones desordenadas, aunque otros umbrales como 0.3 o 0.4 también se utilizan en la literatura (Figura 7.4).

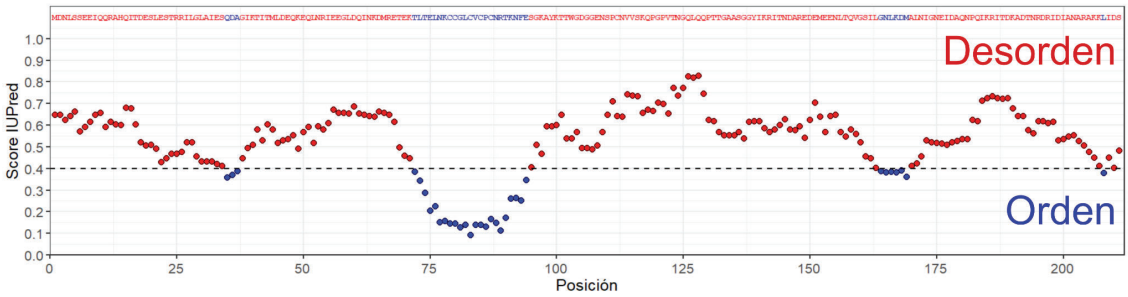


Figura 7.4. Visualización del panel de salida de IUPred. Predicción del desorden de la proteína SNP23_HUMAN (UniProtID: O00161) de los residuos 1-211. Se observa el score de IUPred para cada posición de la secuencia (abajo) y símbolo del aminoácido en esa posición (arriba). La línea horizontal punteada señala el umbral de 0.4 que divide residuos con bajo grado de desorden (color azul) de residuos con alto grado de desorden (color rojo).

A partir de las secuencias completas en formato FASTA de las proteínas presentes en las listas de *hits* de Rb y p107 se calculó el score IUPred por posición, utilizando el parámetro "long" que calcula la probabilidad de contribución energética en una ventana de 100 residuos consecutivos. A partir de los valores obtenidos, para cada péptido *hit* de 16 residuos, utilizando las posiciones de inicio y fin se calculó un valor promedio, desvío estándar y porcentaje de residuos desordenados utilizando un umbral de 0,4.

Desventajas del algoritmo

Dado que IUPred involucra el contexto de la secuencia para calcular el score final, regiones desordenadas en un contexto ordenado son puntuadas con un score bajo de IUPred. Durante la construcción de la biblioteca HD2 para las proteínas transmembrana, los investigadores seleccionaron regiones desordenadas localizadas entre los pasos transmembrana en base a las anotaciones obtenidas de la base de datos Uniprot [40] para incluir la mayor cantidad de interactores posibles en el espacio de búsqueda. Las proteínas transmembrana son ejemplos donde, tanto los segmentos ordenados insertos en la membrana como los loops desordenados que los conectan, son predichos con valores bajos de IUPred (Figura 7.5).

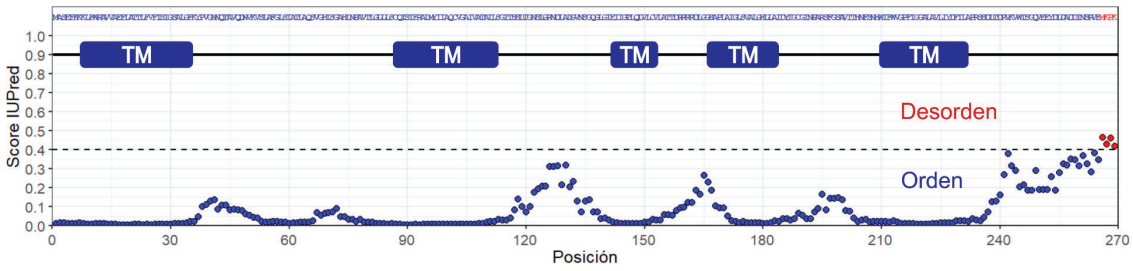


Figura 7.5. Visualización del panel de IUPred de Aquaporina-1. Predicción del desorden de la proteína A0A024RA31_HUMAN de 270 residuos, cuya estructura incluye regiones transmembrana estructuradas y loops desordenados. Los recuadros azules señalan las regiones transmembrana ("TM") de acuerdo a su secuencia en Uniprot (arriba). En azul se muestran las posiciones con un score de IUPred bajo, indicando regiones ordenadas y en rojo, posiciones desordenadas. Debido al proceso de "suavizado" de IUPred, las regiones desordenadas obtienen un score bajo.

7.5.3. Detección de dominios Pfam

La base de datos Pfam reúne información de familias de proteínas representadas por alineamientos múltiples de secuencia y modelos ocultos de Markov (HMM) [56]. La mayoría de los dominios representados en Pfam corresponde a dominios globulares de estructura conservada evolutivamente. Por ejemplo, el subdominio A y el subdominio B de las proteínas *pocket* son cada uno un dominio Pfam (RB_A: PF01858, RB_B: PF01857). Sin embargo, unas pocas familias Pfam corresponden a regiones o proteínas desordenadas que se encuentran altamente conservadas como por ejemplo el dominio de la proteína altamente desordenada p27 Kip1 (CDI: PF02234).

Utilizando los perfiles de la base de datos de Pfam se identificaron para cada proteína con al menos un péptido *hit* en el ensayo de Prop-PD, los dominios Pfam. Luego se indicó para cada péptido *hit* el número de residuos compartidos con dominios Pfam y el dominio Pfam correspondiente.

Cuando un péptido comparte más de ocho residuos con un dominio Pfam, es probable que forme parte de un dominio globular. Dado que la longitud mínima de los SLiMs en este estudio es de cinco residuos, y para ser más inclusivos, se consideró que un péptido no está solapado con un dominio Pfam si comparte ocho o menos residuos. Esto se debe a que estos péptidos podrían estar ubicados en regiones desordenadas adyacentes a un dominio globular.

7.6. Estabilidad energética de péptidos *hit*

FoldX es un campo de fuerza basado en estructura que estima a partir de un complejo proteico entre una proteína globular y un péptido, de manera similar a un modelado del péptido mutado, los cambios en la energía libre de Gibbs (ΔΔG) que ocurren al mutar cada posición del péptido por los 20 aminoácidos posibles [61]. El algoritmo va modificando los aminoácidos de a uno sin alterar la estructura de la cadena lateral o *backbone* del resto del péptido. Como resultado, se obtiene una matriz donde para cada posición del péptido existe un valor de $\Delta\Delta G$. Estos valores luego son normalizados a los residuos del péptido que se encuentran en el complejo, que adopta un valor final de $\Delta\Delta G$ de cero.

7.6.1. Matrices FoldX consideradas en el análisis

Para estimar la estabilidad energética de los péptidos que fueron *hit* de Rb y p107 en el ensayo ProP-PD, se utilizaron matrices de FoldX disponibles en el laboratorio realizadas con FoldX 5.0.

Para el SLiM LxCxE se utilizaron las matrices calculadas a partir de las estructuras resueltas del complejo formado por el dominio *pocket* de Rb y el péptido de 9 residuos de la proteína viral E7 (PDB: 1GUX) [5] (Tabla S9, Anexo) conteniendo el SLiM LxCxE (Figura 7.6) y el complejo formado por el dominio *pocket* de p107 con el péptido de 13 residuos de la proteína LIN52, conteniendo el SLiM LxSxE (PDB: 4YOS) [22] (Tabla S10, Anexo) .

Para el SLiM E2F se utilizaron las matrices calculadas a partir del complejo formado por el dominio *pocket* de Rb con el péptido de 10 residuos de la proteína viral E1A (PDB: 2R7G [33] (Tabla S11, Anexo) y del complejo formado por el dominio *pocket* de Rb con el péptido de 18 residuos de la proteína E2F2 celular (PDB: 1N4M [38] (Tabla S12, Anexo), ambas conteniendo el SLiM E2F.

Las matrices se modificaron de la siguiente manera: en la matriz de 4YOS, se eliminaron los últimos cuatro residuos del péptido cristalizado que no forman parte del SLiM, obteniendo una matriz de 9 residuos comparable al largo de la matriz de 1GUX. En la matriz de 1N4M se eliminaron los primeros nueve residuos del péptido cristalizado que no forman parte del SLiM definido en ELM, obteniendo una matriz de 9 residuos, similar al largo de la matriz de 2R7G.

Matriz FoldX de 1GUX																			
	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
D	0,176	0,339	0,206	0,000	0,148	0,236	0,043	0,193	0,353	0,271	0,294	0,218	-0,031	0,562	0,246	0,230	0,222	0,239	0,331	0,229
L	3,875	3,248	2,952	4,536	3,126	2,534	4,152	5,194	4,303	2,406	0,000	3,633	0,392	0,136	3,197	4,984	3,811	6,093	5,109	2,511
Y	0,914	0,793	1,028	1,054	0,814	0,771	0,755	1,200	0,554	0,126	-0,408	0,276	-0,435	-0,481	1,096	1,192	0,943	-0,281	0,000	0,419
C	0,110	7,258	1,109	0,692	0,000	3,988	2,810	1,699	25,975	6,702	7,664	6,427	2,102	30,966	1,773	1,583	1,289	32,170	32,452	0,516
Y	0,013	0,444	0,063	-0,186	0,025	0,187	-0,171	0,004	0,048	0,184	0,011	0,066	-0,012	-0,017	0,037	-0,019	0,139	0,193	0,000	0,195
E	3,394	4,351	3,254	3,197	2,851	3,198	0,000	4,067	4,638	2,141	2,055	4,106	3,262	4,518	5,283	3,546	3,552	4,485	4,562	2,185
Q	0,114	0,191	0,261	-0,037	0,078	0,000	-0,031	-0,146	0,179	0,040	0,067	0,239	0,035	-0,077	0,108	0,107	0,045	-0,046	-0,016	0,032
L	3,026	5,242	2,411	1,934	2,010	2,911	3,117	3,620	12,105	1,642	0,000	3,488	-0,544	2,037	3,226	3,292	2,239	9,201	4,970	2,144
N	-0,178	0,184	0,000	-0,293	-0,083	0,005	-0,178	-0,563	0,117	0,130	0,186	0,653	-0,017	0,059	3,578	-0,238	-0,155	-0,006	0,043	0,134

Figura 7.6. Matriz FoldX basada en la estructura 1GUX. Matriz provista por el laboratorio de trabajo basada en el complejo formado por el dominio *pocket* de Rb con el péptido de 9 residuos de la proteína viral E7, conteniendo al SLiM LxCxE. En la primera línea se observa el nombre abreviado de cada aminoácido y debajo el valor de la variación de energía libre de Gibbs (ΔΔG) que resulta de sustituir el residuo del péptido cristalizado en la estructura 1GUX por cualquiera de ellos, en kcal. A la izquierda de cada tabla, se observa el símbolo de cada residuo del péptido cristalizado.

7.6.2. Conjunto de datos comparativo

Interactores conocidos de las proteínas *pocket*. Para evaluar la capacidad de FoldX en la detección de SLiMs LxCxE y E2F, se tomaron datos de 50 interactores validados en ELM [36] en febrero 2024 para Rb y p107 conteniendo variantes de los SLiMs LxCxE, LxSxE y E2F, y de siete variantes de la proteína viral E7 del Papilomavirus humano (HPV) testeados experimentalmente en el laboratorio de trabajo. De los 50 interactores, un total de 49 son verdaderos positivos (True Positives, TP) de la proteína *pocket* Rb (Tablas S13 y S15, Anexo) y 37 de p107 (Tabla 7.4) (Tablas S14 y S16, Anexo). Todos los péptidos se diseñaron de manera que tuvieran misma posición de inicio del SLiM y 18 residuos de largo.

De acuerdo a las secuencias recolectadas, se redefinieron las expresiones regulares para el

SLiM LxCxE/LxSxE como:

[IL].[CA].[DE]

[IL].S.[DE]

donde la primera, tercera y quinta son posiciones fijas del SLiM, y también para el SLiM E2F, como:

[IVLMA].[NQDEST][ILFMYWH][ILFMYWH]

donde la primera, tercera, cuarta y quinta son posiciones fijas del SLiM. En las últimas tres posiciones del SLiM, se incluyeron residuos que no se encuentran en la definición de ELM para ser más abarcativos en la búsqueda, dado que muchos de los *hits* presentaron variantes del SLiM E2F.

Tabla 7.4. Interactores TP del conjunto de datos comparativo.

	Rb	**Rb y p107**	**p107**
LxCxE#	10	32	1
E2F*	3	4	-
Total**	**LxCxE = 42** **E2F = 7**	**-**	**LxCxE = 33** **E2F = 4**

#Datos recolectados de 42 interactores de Rb que presentan el SLiM LxCxE (Tabla suplementaria S13),

* Datos recolectado de siete interactores con el SLiM E2F (Tabla suplementaria S15)

** En total son 33 interactores de p107 con el SLiM LxCxE (Tabla suplementaria S14); y cuatro interactores con el SLiM E2F (Tabla suplementaria S16) .

La lista de verdaderos negativos ('True Negative', TN), incluye a los 50 péptidos TP cuya secuencia fue mutada en las posiciones fijas del SLiM por alaninas (A).

La capacidad de matrices FoldX de distinguir péptidos "estables" (TP) de "inestables" (TN) se evaluó con las métricas de *recall* y especificidad de las mismas. El porcentaje de *recall* se calculó de la siguiente manera:

$$Recall = \frac{TP}{TP\ total} * 100$$

donde *TP* es el número de péptidos por debajo de un umbral definido y *TP total* es el número total péptidos TP del conjunto de datos comparativo y su valor puede variar entre 0 y 100%.

El valor de especificidad se calculó de la siguiente manera:

$$Especificidad = \frac{TN}{TN+FP}$$

donde *TN* es el número de péptidos por encima del umbral definido y *TN+FP* es el número de péptidos por encima y por debajo del umbral, o bien el número total de péptidos TN del conjunto de datos comparativos y su valor puede ir de 0 hasta 1.

Péptidos *hit* de ProP-PD testeados experimentalmente. Se incluyó además un conjunto de 62 péptidos *hit* de p107 y Rb obtenidos en el ensayo ProP-PD que el laboratorio ensayó experimentalmente. Para Rb se adicionó el péptido *hit* LIN52 con SLiM LxSxE, que es un interactor validado de p107 pero no de Rb, obteniendo un total de 63 péptidos ensayados experimentalmente por

el laboratorio para esta proteína. Éstos fueron validados utilizando al menos una de las siguientes técnicas *in vitro* de interacción proteína-proteína:

1. *Pull down* con péptidos acoplados a resina (PD/COUP),
2. Cromatografía de exclusión molecular (Size Exclusion Chromatography, SEC) y
3. Ensayo AlphaScreen (AS)

Se descartaron del análisis 21 péptidos en los que no se detectaron las expresiones regulares para los SLiMs LxCxE y E2F y para el caso de Rb, dos péptidos conteniendo el SLiM LxCxE de los que se obtuvieron resultados de mala calidad en el ensayo.

Se obtuvieron finalmente para Rb un total de 40 *hits* ensayados experimentalmente que fueron incluidos en este análisis de los cuales:

- 26 contienen variantes del SLiM LxCxE (Tabla S17, Anexo) y
- 14 contienen variantes del SLiM E2F (Tabla S19, Anexo).

Para el caso de p107, se descartaron 5 péptidos conteniendo el SLiM LxCxE y 2 péptidos conteniendo el SLiM E2F que mostraron una mala calidad en el ensayo con esta proteína *pocket*, obteniéndose en total 34 péptidos *hit* ensayados experimentalmente e incluidos en este análisis, de los cuales:

- 22 contienen variantes del SLiM LxCxE (Tabla S18, Anexo) y
- 12 contienen variantes del SLiM E2F (Tabla S20, Anexo).

Los péptidos fueron clasificados cualitativamente en tres categorías de fuerza de interacción de acuerdo a los resultados de los ensayos de validación experimental llevados a cabo en el laboratorio: *Strong positive* o SP, péptidos de alta afinidad de interacción, *Weak positive* o W, péptidos de afinidad de interacción débil, o bien *Negative* o N, péptidos sin interacción con las proteínas *pocket* (Tabla 7.5). El criterio fue inclusivo, definiendo a un péptido como SP si fue clasificado como SP en al menos una de las técnicas mencionadas anteriormente pero no necesariamente en todas ellas.

Tabla 7.5. Detalle de péptidos *hit* de ProP-PD testeados experimentalmente como parte del conjunto de datos comparativo.

	Rb		p107	
	LxCxE	**E2F**	**LxCxE**	**E2F**
Strong Positive (SP)[#]	12	3	6	2
Weak Positive (W)[#]	9	9	10	3
Negative (N)[#]	5	2	6	7
TOTAL	**26**	**14**	**22**	**12**

[#]Strong Positive (SP): Péptidos de alta afinidad de unión. Weak positive (W):Péptidos de afinidad de interacción débil (Weak positive binder, "W"). Negative (N): Péptidos sin interacción.

Se emplearon las métricas establecidas de *recall* y especificidad para evaluar la capacidad de detección de cada matriz en este conjunto de datos y establecer puntos de corte.

7.6.3. Programa desarrollado de escaneo de secuencias

Las matrices FoldX permiten comparar si un péptido resulta mayor, menor o igual en términos de estabilidad energética que los péptidos identificados en los PDB utilizados. Por lo tanto, se desarrolló un programa de escaneo de secuencias que permitió calcular un valor de $\Delta\Delta G$ para cada péptido *hit* escaneado.

El programa alinea la primera columna de una matriz FoldX con la primera posición del péptido de 16 residuos, luego con la segunda y así sucesivamente avanzando de a una posición, generando subsecuencias de tantos residuos de largo como columnas de la matriz: 1GUX es de 8 o 9 columnas, la matriz 2R7G es de 10, 6 o 5 columnas y 1N4M es de 9, 6 o 5 columnas (ver Sección 7.6.4). A medida que avanza en las posiciones del péptido, calcula el valor final de FoldX como la sumatoria de la variación de energía libre de Gibbs ($\Delta\Delta G$) según la identidad y posición de cada residuo. Hacia el final del péptido, se generan subsecuencias más cortas con una longitud mínima de cinco residuos, que equivale al largo de los residuos centrales de los SLiMs LxCxE y E2F (Figura 7.7).

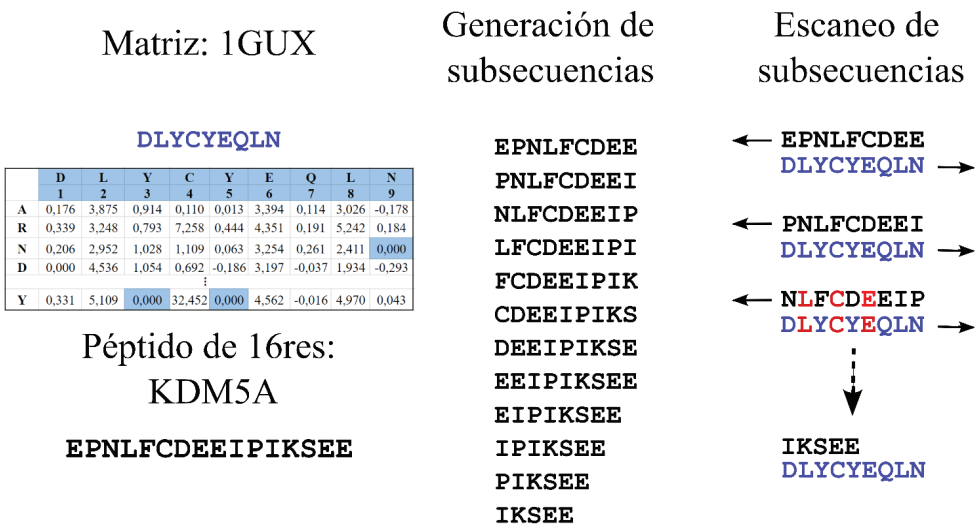


Figura 7.7. Esquema de escaneo de péptidos con matriz 1GUX. El programa desarrollado para escanear secuencias de péptidos utilizando matrices FoldX toma como input una matriz FoldX y un péptido de 16 residuos (panel izquierdo). Se genera inicialmente una subsecuencia del péptido de tantos residuos como largo de la matriz hasta la última subsecuencia de cinco residuos final (panel medio). A medida que avanza posición a posición escaneando la secuencia, calcula el valor de FoldX final como la sumatoria del valor de ΔΔG con el que la matriz penaliza la sustitución en esa posición (panel derecho). En el ejemplo se esquematiza la matriz 1GUX y al péptido de la proteína KDM5A.

Además, se incluyó en el programa la detección de expresiones regulares en las subsecuencias generadas con el fin de identificar el valor de FoldX para subsecuencias que posean el SLiM.

El valor de FoldX para cada subsecuencia puede ser negativo indicando mayor estabilidad energética que el péptido patrón del PDB, positivo indicando menor estabilidad energética que el péptido patrón del PDB, o bien 0 en secuencias idénticas al péptido del PDB. Para cada péptido, se identificaron además dos valores mínimos de FoldX.

7.6.4. Criterios de selección y modificación de matrices FoldX

Para evaluar la capacidad de las matrices en la detección de subsecuencias energéticamente estables que además contienen las expresiones regulares, se realizó un análisis de benchmarking que incluye a los péptidos del conjunto de datos de interactores conocidos y a los péptidos que fueron *hit* en el ensayo ProP-PD que fueron testeados experimentalmente.

Se utilizaron las métricas de *recall* y especificidad para comparar la estabilidad de péptidos de interactores conocidos o testeados según sus valores y para determinar los puntos de corte que indican la estabilidad relativa de un péptido. Los umbrales definidos se aplicaron posteriormente para evaluar energéticamente la lista completa de péptidos *hit* de ProP-PD.

Por otro lado, se estableció la métrica de porcentaje de recuperación de péptidos. A diferencia de la métrica de *recall* utilizada en secciones anteriores donde se calcula el porcentaje en base a cuántos péptidos fueron recuperados como *hit* en el ensayo ProP-PD en relación con los péptidos

presentes en la biblioteca HD2, en el porcentaje de *recall* se cuenta el número de *hits* que se encuentran por debajo del umbral establecido para cada SLiM, en relación con el total de péptidos dentro del grupo con SLiM LxCxE, E2F o sin SLiM detectado, de la siguiente manera:

$$Recall = \frac{Npep}{Ntotal} * 100$$

donde *Npep* corresponde al número total de péptidos *hit* por debajo del umbral considerado y *Ntotal* corresponde al total de péptidos con o sin SLiM detectado en cada caso.

Matriz 4YOS. Se consideró la matriz 4YOS para analizar péptidos en los que se detecta el SLiM LxCxE (Figura 7.8).

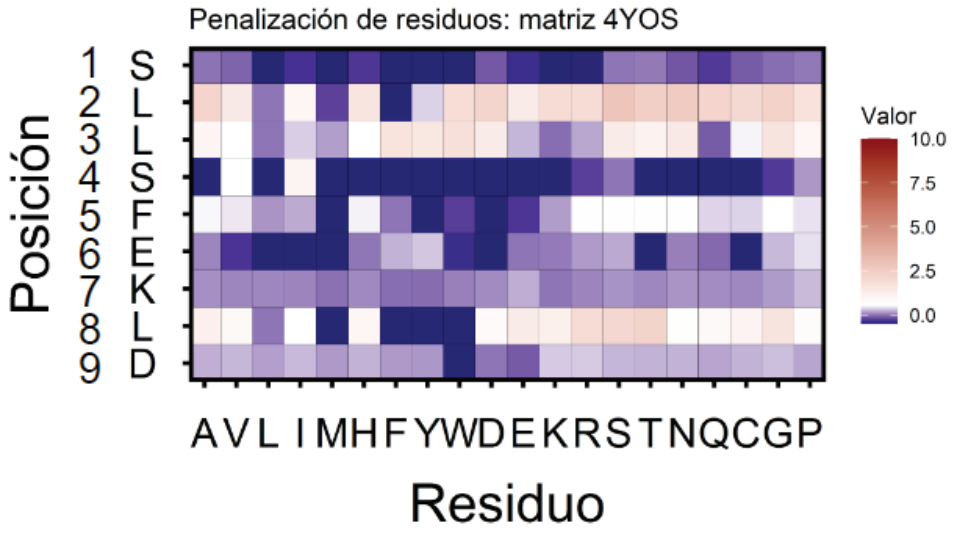


Figura 7.8. Penalización de residuos de la matriz FoldX de 4YOS. Heatmap construido a partir de los residuos que figuran en el PDB 4YOS (eje de ordenadas). Las posiciones del péptido se encuentran numeradas desde serina (S) en la posición uno hasta aspártico (D) en la posición nueve. En el eje de abscisas se observan los 20 aminoácidos posibles para sustituir en el péptido y en colores que van desde el azul -sustituciones favorables- al rojo-sustituciones desfavorables- se simboliza la penalización de cada aminoácido en los residuos de la matriz.

Los residuos centrales del SLiM LxSxE, ubicados en las posiciones dos, cuatro y seis de esta matriz son los encargados de mediar la interacción con el dominio pocket. Si bien en la segunda posición se encuentran penalizados la mayoría de las sustituciones, las posiciones cuatro y seis admiten todos los residuos, resultando muy permisivas.

Por otro lado, la octava posición presenta valores de penalización más altos en comparación con los residuos del núcleo del SLiM en las posiciones dos, cuatro y seis, lo que sugiere que esta posición, que involucra residuos hidrofóbicos, es crucial para mediar la unión.

Sin embargo, las penalizaciones en las posiciones centrales del SLiM, que entran en contacto directo con la superficie del bolsillo, no condicen con la evidencia experimental de determinantes de

afinidad de los SLiMs LxCxE y LxSxE conocidos, por lo que fue descartada como herramienta de priorización de péptidos [34].

Matriz 1GUX. La matriz FoldX 1GUX disponible en el laboratorio de trabajo, fue considerada para evaluar la estabilidad de péptidos *hit* del ensayo ProP-PD (Figura 7.9).

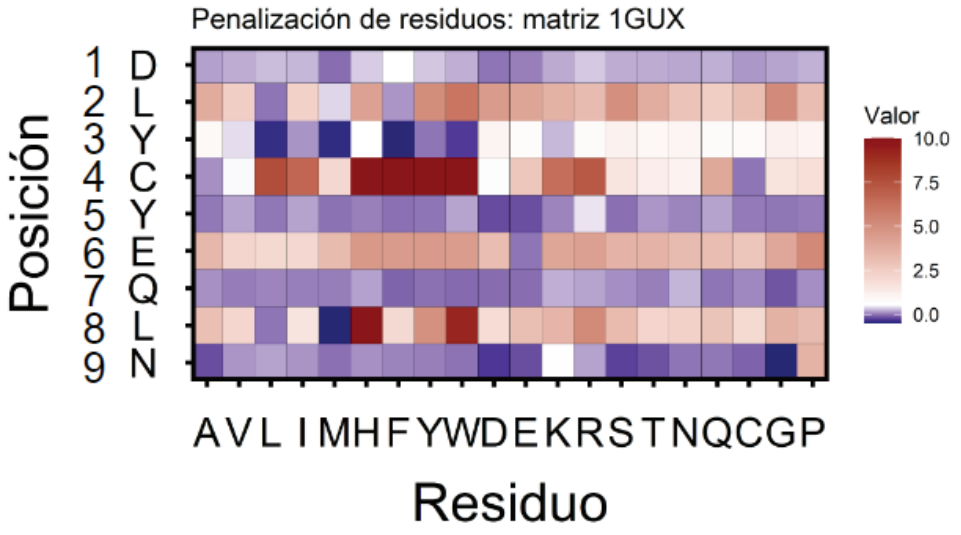


Figura 7.9. Penalización de residuos de la matriz FoldX de 1GUX. Heatmap construido a partir de los residuos que figuran en el PDB 1GUX (eje de ordenadas). Las posiciones del péptido se encuentran numeradas desde el aspártico (D) en la posición uno hasta asparagina (N) en la posición nueve. En el eje de abscisas se observan los 20 aminoácidos posibles para sustituir en el péptido y en colores que van desde el azul -sustituciones favorables- al rojo-sustituciones desfavorables- se simboliza la penalización de cada aminoácido en los residuos de la matriz.

En la primera posición del péptido, FoldX no penaliza fuertemente ningún aminoácido y por lo tanto todos son permitidos. La segunda, cuarta y sexta posición son los residuos centrales del SLiM y se encuentran altamente penalizados los residuos aromáticos F, Y o W e H en la cuarta posición, mientras que en la segunda y sexta, se admite la sustitución de muy pocos residuos.

Las posiciones tres y cinco, corresponden a las posiciones variables del centro del SLiM y los cambios de aminoácidos se encuentran poco penalizados.

Residuos hidrofóbicos en las posiciones +2 o +3 con respecto al ***core*** del SLiM también participan en la interacción con el dominio *pocket* [34,50]. En la matriz 1GUX, estas posiciones corresponden a las ocho y nueve. La penalidad en la octava posición es menor para los residuos hidrofóbicos, mientras que la novena posición admite una variedad de residuos sin distinguir entre hidrofóbicos y no hidrofóbicos. Aunque la presencia de un residuo hidrofóbico en las posiciones +2 o +3 favorece la interacción, la matriz no permite distinguir entre péptidos con residuos no hidrofóbicos en la posición +3 y aquellos que sí los presentan.

Tomando en consideración que en la primera posición se penaliza con valores cercanos a cero

la sustitución de todos los aminoácidos, se decidió generar una variante de 1GUX, eliminándola de la matriz de escaneo. A la variante que incluye los residuos de las posiciones uno a nueve, se la llamó 1GUX_9 (DLYCYEQLN) y a la variante con los residuos de las posiciones dos a nueve, se la llamó 1GUX_8 (LYCYEQLN).

Matriz 2R7G. Para el análisis de estabilidad energética de péptidos *hit* conteniendo el SLiM E2F, se utilizó en primera instancia la matriz FoldX del PDB 2R7G [33] (Figura 7.10).

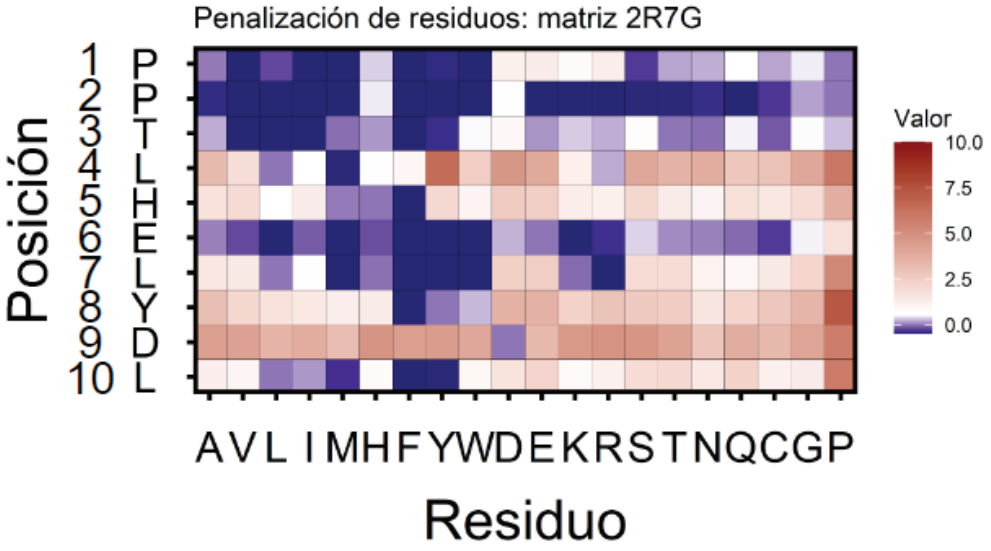


Figura 7.10. Penalización de residuos de la matriz FoldX de 2R7G. Heatmap construido a partir de los residuos que figuran en el PDB 2R7G [33] (eje de ordenadas). Las posiciones del péptido se encuentran numeradas desde la prolina (P) en la posición uno hasta lisina (L) en la posición diez. En el eje de abscisas se observan los 20 aminoácidos posibles para sustituir en el péptido y en colores que van desde el azul -sustituciones favorables- al rojo-sustituciones desfavorables- se simboliza la penalización de cada aminoácido en los residuos de la matriz.

Las primeras tres posiciones de la matriz 2R7G se encuentran en un rango de valores de penalización cercanos a cero o bien favorecen algunos residuos hidrofóbicos, aunque no hay evidencia experimental o variantes naturales que avalen esta preferencia [33].

El SLiM E2F comienza en la cuarta posición, que admite preferentemente residuos hidrofóbicos. La quinta posición es variable en el SLiM. La sexta y séptima posición penalizan con valores cercanos a cero o bien admiten la sustitución de residuos, mientras que la posición ocho admite únicamente residuos aromáticos o hidrofóbicos. La novena posición penaliza todos los residuos, incrementando el valor de estabilidad energética, al contrario de la décima posición que admite todas las sustituciones menos la prolina y ninguna de las dos forma parte de los residuos centrales del SLiM.

Se generaron tres variantes de esta matriz, teniendo en cuenta que las primeras tres posiciones y las dos últimas no forman parte del centro del SLiM. La primera de las variantes es la matriz

original del péptido que incluye a los residuos de las posiciones uno a diez, a la que se llamó 2R7G_10 (PPTLHELYDL). Luego, debido a la falta de evidencia experimental y al alto rango de penalización se utilizó una matriz que incluye al péptido de la cuarta a la novena posición a la que se llamó 2R7G_6 (LHELYD) y otra del péptido de la cuarta a la octava posición, denominada 2R7G_5 (LHELY).

Matriz 1N4M. Por último se consideró además de 2R7G, la matriz FoldX 1N4M para evaluar la estabilidad energética de péptidos *hit* con SLiM E2F [38] (Figura 7.11).

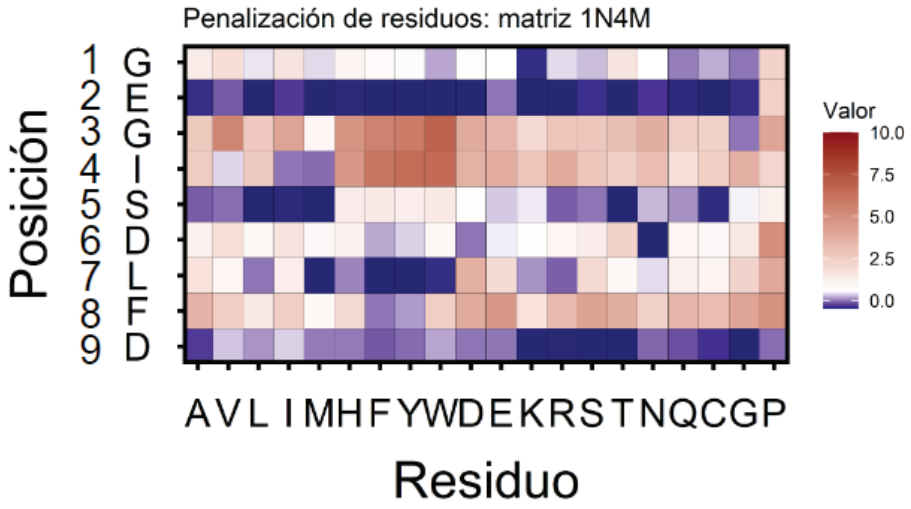


Figura 7.11. Penalización de residuos de la matriz FoldX de 1N4M. Heatmap construido a partir de los residuos que figuran en el PDB 1N4M (eje de ordenadas). Las posiciones del péptido se encuentran numeradas desde la glicina (G) en la posición uno hasta el aspártico (D) en la posición nueve. En el eje de abscisas se observan los 20 aminoácidos posibles para sustituir en el péptido y en colores que van desde el azul -sustituciones favorables- al rojo-sustituciones desfavorables- se simboliza la penalización de cada aminoácido en los residuos de la matriz.

En el mapa de la matriz 1N4M se observó que las primeras dos posiciones admiten la sustitución de todos los residuos, menos la prolina, mientras que la tercera posición penaliza 18 de los 20 aminoácidos incrementando el valor de estabilidad energética.

Los residuos centrales del SLiM abarcan desde la cuarta hasta la octava posición siendo éstas dos las que más sustituciones penalizan, en comparación a la quinta, sexta y séptima. La novena posición favorece la unión entre el péptido [38] y la proteína *pocket*, y la matriz no refleja esto dado que admite cualquier sustitución.

Teniendo en cuenta las observaciones de las primeras tres posiciones, que no forman parte del centro del SLiM, y de la novena posición, se utilizaron tres variantes de la matriz. La primera incluye al péptido de nueve residuos que incluye a los aminoácidos de la posición uno a nueve, llamada 1N4M_9 (GEGISDLFD), otra que incluye a los residuos de las posiciones cuatro a nueve denominada 1N4M_6 (ISDLFD) y por último, una que incluye los residuos de las posiciones cuatro a ocho, 1N4M_5 (ISDLF).

El presente manuscrito corresponde a la versión final de la Tesis de Licenciatura realizada por Carla Lorenze, bajo la dirección de la Dra. Lucia B. Chemes y la co-dirección de la Dra. Juliana Glavina. Esta versión incluye correcciones y sugerencias realizadas por el jurado durante la defensa de la misma.

Alumna Tesista: Carla Lorenze

Directora: Dra. Lucia B. Chemes

Co-Directora: Juliana Glavina

Capítulo 8: Referencias

1. van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, et al. Classification of intrinsically disordered regions and proteins. Chem Rev. 2014;114: 6589–6631.

2. Van Roey K, Uyar B, Weatheritt RJ, Dinkel H, Seiler M, Budd A, et al. Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. Chem Rev. 2014;114: 6733–6778.

3. Syed RS, Reid SW, Li C, Cheetham JC, Aoki KH, Liu B, et al. Efficiency of signalling through cytokine receptors depends critically on receptor orientation. Nature. 1998;395: 511–516.

4. Russo AA, Jeffrey PD, Patten AK, Massagué J, Pavletich NP. Crystal structure of the p27Kip1 cyclin-dependent-kinase inhibitor bound to the cyclin A-Cdk2 complex. Nature. 1996;382: 325–331.

5. Lee JO, Russo AA, Pavletich NP. Structure of the retinoblastoma tumour-suppressor pocket domain bound to a peptide from HPV E7. Nature. 1998;391: 859–865.

6. Babu MM, Kriwacki RW, Pappu RV. Structural biology. Versatility from protein disorder. Science. 2012;337: 1460–1461.

7. Uversky VN. Unusual biophysics of intrinsically disordered proteins. Biochim Biophys Acta. 2013;1834: 932–951.

8. Brown CJ, Johnson AK, Dunker AK, Daughdrill GW. Evolution and disorder. Curr Opin Struct Biol. 2011;21: 441–446.

9. Anfinsen CB, Haber E, Sela M, White FH Jr. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc Natl Acad Sci U S A. 1961;47: 1309–1314.

10. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol. 2005;6: 197–208.

11. Huang A, Stultz CM. Finding order within disorder: elucidating the structure of proteins associated with neurodegenerative disease. Future Med Chem. 2009;1: 467–482.

12. Oldfield CJ, Dunker AK. Intrinsically disordered proteins and intrinsically disordered protein regions. Annu Rev Biochem. 2014;83: 553–584.

13. Zhang J, Fan J-S, Li S, Yang Y, Sun P, Zhu Q, et al. Structural basis of DNA binding to human YB-1 cold shock domain regulated by phosphorylation. Nucleic Acids Res. 2020;48: 9361–9371.

14. Song J, Lee MS, Carlberg I, Vener AV, Markley JL. Micelle-induced folding of spinach thylakoid soluble phosphoprotein of 9 kDa and its functional implications. Biochemistry. 2006;45: 15633–15643.

15. Dick FA, Rubin SM. Molecular mechanisms underlying RB protein function. Nature Reviews Molecular Cell Biology. 2013. pp. 297–306. doi:10.1038/nrm3567

16. Forman-Kay JD, Mittag T. From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. Structure. 2013;21: 1492–1499.

17. Davey NE, Van Roey K, Weatheritt RJ, Toedt G, Uyar B, Altenberg B, et al. Attributes of short linear motifs. Mol Biosyst. 2012;8: 268–281.

18. Dinkel H, Van Roey K, Michael S, Davey NE, Weatheritt RJ, Born D, et al. The eukaryotic linear motif resource ELM: 10 years and counting. Nucleic Acids Res. 2014;42: D259–66.

19. Orndorff KS, Veltri EJ, Hoitsma NM, Williams IL, Hall I, Jaworski GE, et al. Structural Basis for the Interaction Between Yeast Chromatin Assembly Factor 1 and Proliferating Cell Nuclear Antigen. J Mol Biol. 2024;436: 168695.

20. Prestel A, Wichmann N, Martins JM, Marabini R, Kassem N, Broendum SS, et al. The PCNA interaction motifs revisited: thinking outside the PIP-box. Cell Mol Life Sci. 2019;76: 4923–4943.

21. Classon M, Harlow E. The retinoblastoma tumour suppressor in development and cancer. Nat Rev Cancer. 2002;2: 910–917.

22. Guiley KZ, Liban TJ, Felthousen JG, Ramanan P, Litovchick L, Rubin SM. Structural mechanisms of DREAM complex assembly and regulation. Genes Dev. 2015;29: 961–974.

23. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022;50: D439–D444.

24. González-Foutel NS, Glavina J, Borcherds WM, Safranchik M, Barrera-Vilarmau S, Sagar A, et al. Conformational buffering underlies functional selection in intrinsically disordered protein regions. Nat Struct Mol Biol. 2022;29: 781–790.

25. Venkadakrishnan VB, Yamada Y, Weng K, Idahor O, Beltran H. Significance of RB Loss in Unlocking Phenotypic Plasticity in Advanced Cancers. Mol Cancer Res. 2023;21: 497–510.

26. Classon M, Dyson N. p107 and p130: versatile proteins with interesting pockets. Exp Cell Res. 2001;264: 135–147.

27. Coschi CH, Martens AL, Ritchie K, Francis SM, Chakrabarti S, Berube NG, et al. Mitotic chromosome condensation mediated by the retinoblastoma protein is tumor-suppressive. Genes Dev. 2010;24: 1351–1363.

28. Burkhart DL, Wirt SE, Zmoos A-F, Kareta MS, Sage J. Tandem E2F binding sites in the promoter of the p107 cell cycle regulator control p107 expression and its cellular functions. PLoS Genet. 2010;6: e1001003.

29. Thompson K. Programming Techniques: Regular expression search algorithm. Commun ACM. 1968;11: 419–422.

30. * Standards IEE. 1003.1-2001 Posix Pt.1:Open Group Tech Std Issu: Open Group Technical Standard. IEEE; 2001.

31. Edwards RJ, Palopoli N. Computational prediction of short linear motifs from protein sequences. Methods Mol Biol. 2015;1268: 89–141.

32. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14: 1188–1190.

33. Liu X, Marmorstein R. Structure of the retinoblastoma protein bound to adenovirus E1A reveals the molecular basis for viral oncoprotein inactivation of a tumor suppressor. Genes Dev. 2007;21: 2711–2716.

34. Palopoli N, González Foutel NS, Gibson TJ, Chemes LB. Short linear motif core and flanking regions modulate retinoblastoma protein binding affinity and specificity. Protein Eng Des Sel. 2018;31: 69–77.

35. Glavina J, Román EA, Espada R, de Prat-Gay G, Chemes LB, Sánchez IE. Interplay between sequence, structure and linear motifs in the adenovirus E1A hub protein. Virology. 2018;525: 117–131.

36. Kumar M, Gouw M, Michael S, Sámano-Sánchez H, Pancsa R, Glavina J, et al. ELM-the eukaryotic linear motif resource in 2020. Nucleic Acids Res. 2020;48: D296–D306.

37. Palopoli N, González Foutel NS, Gibson TJ, Chemes LB. Short linear motif core and flanking regions modulate retinoblastoma protein binding affinity and specificity. Protein Eng Des Sel. 2018;31: 69–77.

38. Lee C, Chang JH, Lee HS, Cho Y. Structural basis for the recognition of the E2F transactivation domain by the retinoblastoma tumor suppressor. Genes Dev. 2002;16: 3199–3212.

39. Tompa P, Davey NE, Gibson TJ, Babu MM. A million peptide motifs for the molecular biologist. Mol Cell. 2014;55: 161–169.

40. Benz C, Ali M, Krystkowiak I, Simonetti L, Sayadi A, Mihalic F, et al. Proteome-scale mapping of binding sites in the unstructured regions of the human proteome. Mol Syst Biol. 2022;18: e10584.

41. Ivarsson Y, Jemth P. Affinity and specificity of motif-based protein-protein interactions. Curr Opin Struct Biol. 2019;54: 26–33.

42. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, et al. The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 2014;42: D358–63.

43. Davey NE, Seo M-H, Yadav VK, Jeon J, Nim S, Krystkowiak I, et al. Discovery of short linear motif-mediated interactions through phage display of intrinsically disordered regions of the human proteome. FEBS J. 2017;284: 485–498.

44. Ivarsson Y, Arnold R, McLaughlin M, Nim S, Joshi R, Ray D, et al. Large-scale interaction profiling of PDZ domains through proteomic peptide-phage display using human and viral phage peptidomes. Proc Natl Acad Sci U S A. 2014;111: 2542–2547.

45. Wigington CP, Roy J, Damle NP, Yadav VK, Blikstad C, Resch E, et al. Systematic Discovery of Short Linear Motifs Decodes Calcineurin Phosphatase Signaling. Mol Cell. 2020;79: 342–358.e12.

46. Lüchow S, Sundell GN, Ivarsson Y. Identification of PDZ Interactions by Proteomic Peptide Phage Display. Methods Mol Biol. 2021;2256: 41–60.

47. Mihalič F, Benz C, Kassa E, Lindqvist R, Simonetti L, Inturi R, et al. Identification of motif-based interactions between SARS-CoV-2 protein domains and human peptide ligands pinpoint antiviral targets. Nat Commun. 2023;14: 5636.

48. Sanidas I, Morris R, Fella KA, Rumde PH, Boukhali M, Tai EC, et al. A Code of Mono-phosphorylation Modulates the Function of RB. Mol Cell. 2019;73: 985–1000.e6.

49. Kumar M, Michael S, Alvarado-Valverde J, Zeke A, Lazar T, Glavina J, et al. ELM-the Eukaryotic Linear Motif resource-2024 update. Nucleic Acids Res. 2024;52: D442–D455.

50. Putta S, Alvarez L, Lüdtke S, Sehr P, Müller GA, Fernandez SM, et al. Structural basis for tunable affinity and specificity of LxCxE-dependent protein interactions with the retinoblastoma protein family. Structure. 2022;30: 1340–1353.e3.

51. Keskin O, Tuncbag N, Gursoy A. Predicting Protein-Protein Interactions from the Molecular to the Proteome Level. Chem Rev. 2016;116: 4884–4909.

52. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009;37: W202–8.

53. Singh M, Krajewski M, Mikolajka A, Holak TA. Molecular determinants for the complex formation between the retinoblastoma protein and LXCXE sequences. J Biol Chem. 2005;280: 37868–37876.

54. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596: 583–589.

55. Mészáros B, Erdos G, Dosztányi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Res. 2018;46: W329–W337.

56. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021;49: D412–D419.

57. Shrake A, Rupley JA. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J Mol Biol. 1973;79: 351–371.

58. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. Highly accurate protein structure prediction for the human proteome. Nature. 2021;596: 590–596.

59. Pereira J, Simpkin AJ, Hartmann MD, Rigden DJ, Keegan RM, Lupas AN. High-accuracy protein structure prediction in CASP14. Proteins. 2021;89: 1687–1699.

60. Piovesan D, Monzon AM, Tosatto SCE. Intrinsic protein disorder and conditional folding in AlphaFoldDB. Protein Sci. 2022;31: e4466.

61. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucleic Acids Res. 2005;33: W382–8.

62. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28: 235–242.

63. Jiang Y, Mu H, Zhao H. HMBOX1, a member of the homeobox family: current research progress. Cent Eur J Immunol. 2023;48: 63–69.

64. Chen S, Yang Z, Wilkinson AW, Deshpande AJ, Sidoli S, Krajewski K, et al. The PZP Domain of AF10 Senses Unmodified H3K27 to Regulate DOT1L-Mediated Methylation of H3K79. Mol Cell. 2015;60: 319–327.

65. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023;51: D523–D531.

66. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25: 25–29.

67. Huang H, Sidhu SS. Studying binding specificities of peptide recognition modules by high-throughput phage display selections. Methods Mol Biol. 2011;781: 87–97.

68. Ali M, Simonetti L, Ivarsson Y. Screening Intrinsically Disordered Regions for Short Linear Binding Motifs. Methods Mol Biol. 2020;2141: 529–552.

69. Liban TJ, Thwaites MJ, Dick FA, Rubin SM. Structural Conservation and E2F Binding Specificity within the Retinoblastoma Pocket Protein Family. J Mol Biol. 2016;428: 3960–3971.

70. Bailey TL. Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Bipolymers. 1994.

71. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22: 2577–2637.

Capítulo 9: Anexo

Tabla S1. Interactores humanos de las proteínas *pocket* Rb y p107 conteniendo a los SLiMs LxCxE y E2F reportados en la base de datos ELM.

Uniprot ID	Accession	Nombre	SLiM[#]	Secuencia*	Sub-Secuencia**	FoldX***	HD2	Rb	p107	p130
ARI4A_HUMAN	P29374	AT-rich interactive domain-containing protein 4A	LIG_Rb_LxCxE_1	GPET**LVCHE**VDLDDLDEK	LVCHEVDL	2.62	SI	SI	SI	
CCND1_HUMAN	P24385	G1/S-specific cyclin-D1	LIG_Rb_LxCxE_1	MEHQ**LLCCE**VETIRRAYP	LLCCEVET	2.61	SI	SI	SI	
CCND2_HUMAN	P30279	G1/S-specific cyclin-D2	LIG_Rb_LxCxE_1	ME**LLCHE**VDPVRRAVR	LLCHEVDP	5.18	NO	SI	SI	
CCND3_HUMAN	P30281	G1/S-specific cyclin-D3	LIG_Rb_LxCxE_1	ME**LLCCE**GTRHAPRAG	LLCCEGTR	1.89	SI	SI	SI	
EID1_HUMAN	Q9Y6B2	EP300-interacting inhibitor of differentiation 1	LIG_Rb_LxCxE_1	LTEE**LGCDE**IIDRE	LGCDEIID	2.4	SI	SI	SI	
HDAC1_HUMAN	Q13547	Histone deacetylase 1 (HD1)	LIG_Rb_LxCxE_1	SDKR**IACEE**EFSDSEEEG	IACEEEFS	4.92	SI	SI	SI	
HDAC2_HUMAN	Q92769	Histone deacetylase 2 (HD2)	LIG_Rb_LxCxE_1	SDKR**IACDE**EFSDSEDEG	IACDEEFS	4.9	SI	SI		
KDM5A_HUMAN	P29375	Lysine-specific demethylase 5A	LIG_Rb_LxCxE_1	LEPN**LFCDE**EIPIKSEEV	IKSEEV	4.13	SI	SI	SI	
NDC80_HUMAN	O14777	Kinetochore protein NDC80 homolog	LIG_Rb_LxCxE_1	LDYT**IKCYE**SFMSGADSF	IKCYESFM	4.81	NO	SI		
PPR26_HUMAN	Q5T8A7	Protein phosphatase 1 regulatory subunit 26	LIG_Rb_LxCxE_1	TSAE**LMCAE**AILDISKTI	LMCAEAIL	1.52	SI	SI		
PRDM2_HUMAN	Q13029	PR domain zinc finger protein 2	LIG_Rb_LxCxE_1	KEPE**IRCDE**KPEDLLEEP	IRCDEKPE	6.3	SI	SI	SI	
SMCA2_HUMAN	P51531	Probable global transcription activator SNF2L2	LIG_Rb_LxCxE_1	EVER**LTCEE**EEEKIFGRG	LTCEEEEE	3.68	SI	SI	SI	
SMCA4_HUMAN	P51532	Transcription activator BRG1	LIG_Rb_LxCxE_1	EVER**LTCEE**EEEKMFGRG	LTCEEEEE	3.68	SI	SI	SI	
E2F1_HUMAN	Q01094	Transcription factor E2F1	LIG_Rb_pABgroove_1	LDYHFGLEEGEG**IRDLF**D	IRDLF	-0.1	SI	SI	SI	
E2F2_HUMAN	Q14209	Transcription factor E2F2	LIG_Rb_pABgroove_1	DDYLWGLEAGEG**ISDLF**D	ISDLF	0	SI	SI	SI	

Uniprot ID	Accession	Nombre	SLiM[#]	Secuencia*	Sub-Secuencia**	FoldX***	HD2	Rb	p107	p130
E2F3_HUMAN	O00716	Transcription factor E2F3	LIG_Rb_pABgroove_1	EDYLLSLGEEEG**ISDLF**D	ISDLF	0	**SI**	**SI**	**SI**	
E2F4_HUMAN	Q16254	Transcription factor E2F4	LIG_Rb_pABgroove_1	HDYIYNLDESEG**VCDLF**D	VCDLF	-0.04	**SI**	**SI**	**SI**	
E2F5_HUMAN	Q15329	Transcription factor E2F5	LIG_Rb_pABgroove_1	DDYNFNLDDNEG**VCDLF**D	VCDLF	-0.04	**SI**	**SI**		
LIN52_HUMAN	Q52LA3	Protein lin-52 homolog	LIG_RBL1_LxSxE_2	LEAS**LLSFE**KLDRASPDL	LLSFEKLD	1.1	**SI**		**SI**	**SI**

[#]identificador ELM del SLiM; SLiM LxCxE: LIG_Rb_LxCxE_1, SLiM LxSxE: LIG_RBL1_LxSxE_2, SLiM E2F: LIG_Rb_pABgroove_1

*secuencia del péptido conteniendo al SLiM con el centro resaltado [36]

**Sub-secuencia escaneada por la matriz FoldX

***Valor FoldX de matrices 1GUX_8 en los SLiMs LxCxE/LxSxE o 1N4M_5 en los SLiMs E2F

Tabla S2. Afinidades de interacciones proteína-proteína utilizadas en el trabajo.

Proteína pocket	Proteína	Péptido	Secuencia[#]	Afinidad (K_D, µM)	Referencia
Rb	KDM5A_HUMAN	*wild-type*	EPN**L**F**C**D**E**EIPIKSEE	0.6 ± 0.1	[50]
p107	KDM5A_HUMAN	*Wild-type*	EPN**L**F**C**D**E**EIPIKSEE	0.5 ± 0.1	[50]
p107	LIN52	*Wild-type*	LEAS**L**L**S**F**E**KLDRASP	5.9 ± 0.9	[50]
Rb	E7 - HPV	*Wild-type*	QPETTD**L**Y**C**Y**E**QLNDS	0.007	Datos no publicados
Rb	E7 - HPV	C24A	QPETTD**L**Y**A**Y**E**QLNDS	0.435	Datos no publicados
Rb	E7 - HPV	C24S	QPETTD**L**Y**S**Y**E**QLNDS	1.542	Datos no publicados
Rb	E7 - HPV	L28W	QPETTD**L**Y**C**Y**E**Q**W**NDS	0.013	Datos no publicados

[#]Secuencia del péptido donde se destaca la expresión regular del SLiM LxCxE (negro) y mutaciones (rojo).

Interactores conocidos reportados en IntAct (Noviembre 2021)

Tabla S3. Interactores de la proteína Rb reportados en la base de datos de interactores, IntAct.

UniprotID	Accession	Nombre	Tipo de Interacción	Hit ProP-PD*	Proteómica**
CDK4_HUMAN	P11802	Cyclin-dependent kinase 4	Directa		
RT18B_HUMAN	Q9Y676	Small ribosomal subunit protein mS40	Directa		
NTM1A_HUMAN	Q9BV86	N-terminal Xaa-Pro-Lys N-methyltransferase 1	Directa		
SETD7_HUMAN	Q8WTS6	Histone-lysine N-methyltransferase SETD7	Directa		
MCM7_HUMAN	P33993	DNA replication licensing factor MCM7	Directa		
CDK6_HUMAN	Q00534	Cyclin-dependent kinase 6	Directa		SI
CDK2_HUMAN	P24941	Cyclin-dependent kinase 2	Directa		SI
CHK1_HUMAN	O14757	Serine/threonine-protein kinase Chk1	Directa		
MK14_HUMAN	Q16539	Mitogen-activated protein kinase 14	Directa		
EP300_HUMAN	Q09472	Histone acetyltransferase p300	Directa		
PRS6B_HUMAN	P43686	26S proteasome regulatory subunit 6B	Directa		
SIR1_HUMAN	Q96EB6	NAD-dependent protein deacetylase sirtuin-1	Directa		
RB_HUMAN	P06400	Retinoblastoma-associated protein	Directa-Indirecta		
ANM2_HUMAN	P55345	Protein arginine N-methyltransferase 2	Directa-Indirecta		
HDAC1_HUMAN	Q13547	Histone deacetylase 1	Directa-Indirecta		SI
E2F2_HUMAN	Q14209	Transcription factor E2F2	Directa-Indirecta	SI	
FRK_HUMAN	P42685	Tyrosine-protein kinase FRK	Directa-Indirecta		
PP1A_HUMAN	P62136	Serine/threonine-protein phosphatase PP1-alpha catalytic subunit	Directa-Indirecta		SI
CUX1_HUMAN	P39880	Homeobox protein cut-like 1	Directa-Indirecta		
E2F5_HUMAN	Q15329	Transcription factor E2F5	Directa-Indirecta	SI	SI
E2F1_HUMAN	Q01094	Transcription factor E2F1	Directa-Indirecta	SI	SI
NCOA6_HUMAN	Q14686	Nuclear receptor coactivator 6	Directa-Indirecta		
CHK2_HUMAN	O96017	Serine/threonine-protein kinase Chk2	Directa-Indirecta		
UBP7_HUMAN	Q93009	Ubiquitin carboxyl-terminal hydrolase 7	Directa-Indirecta		
E2F4_HUMAN	Q16254	Transcription factor E2F4	Indirecta		SI

UniprotID	Accession	Nombre	Tipo de Interacción	Hit ProP-PD*	Proteómica**
LMNA_HUMAN	P02545	Prelamin-A/C [Cleaved into: Lamin-A/C]	Indirecta		
ECD_HUMAN	O95905	Protein ecdysoneless homolog	Indirecta	SI	
TRMO_HUMAN	Q9BU70	tRNA-N6)-methyltransferase	Indirecta		
CC180_HUMAN	Q9P1Z9	Coiled-coil domain-containing protein 180	Indirecta		
RBG1L_HUMAN	Q5R372	Rab GTPase-activating protein 1-like	Indirecta		
IF16_HUMAN	Q16666	Gamma-interferon-inducible protein 16	Indirecta		
ANS1A_HUMAN	Q92625	Ankyrin repeat and SAM domain-containing protein 1A	Indirecta		
DGKZ_HUMAN	Q13574	Diacylglycerol kinase zeta	Indirecta		
E2F3_HUMAN	O00716	Transcription factor E2F3	Indirecta	SI	SI
NPM_HUMAN	P06748	Nucleophosmin	Indirecta		
CLNK_HUMAN	Q7Z7G1	Cytokine-dependent hematopoietic cell linker	Indirecta		
PA2G4_HUMAN	Q9UQ80	Proliferation-associated protein 2G4	Indirecta		
PURA_HUMAN	Q00577	Transcriptional activator protein Pur-alpha	Indirecta		
AHR_HUMAN	P35869	Aryl hydrocarbon receptor	Indirecta		
CCNC_HUMAN	P24863	Cyclin-C	Indirecta		
UBF1_HUMAN	P17480	Nucleolar transcription factor 1	Indirecta		
FANCC_HUMAN	Q00597	Fanconi anemia group C protein	Indirecta		
HBP1_HUMAN	O60381	HMG box-containing protein 1	Indirecta		
TFDP1_HUMAN	Q14186	Transcription factor Dp-1	Indirecta		SI
COR2A_HUMAN	Q92828	Coronin-2A	Indirecta		
BRE1B_HUMAN	O75150	E3 ubiquitin-protein ligase BRE1B	Indirecta		
GLT12_HUMAN	Q8IXK2	Polypeptide N-acetylgalactosaminyltransferase 12	Indirecta		
CHIO_HUMAN	P52757	Beta-chimaerin	Indirecta		
PCGF3_HUMAN	Q3KNV8	Polycomb group RING finger protein 3	Indirecta		
NFM_HUMAN	P07197	Neurofilament medium polypeptide	Indirecta		
HELZ_HUMAN	P42694	Probable helicase with zinc finger domain	Indirecta		
STX17_HUMAN	P56962	Syntaxin-17	Indirecta		
PAX8_HUMAN	Q06710	Paired box protein Pax-8	Indirecta		

UniprotID	Accession	Nombre	Tipo de Interacción	Hit ProP-PD*	Proteómica**
MO4L2_HUMAN	Q15014	Mortality factor 4-like protein 2	Indirecta		
CTIP_HUMAN	Q99708	DNA endonuclease RBBP8	Indirecta		
PG12A_HUMAN	Q9BZM1	Group XIIA secretory phospholipase A2	Indirecta		
CATL2_HUMAN	O60911	Cathepsin L2	Indirecta		
F16P2_HUMAN	O00757	Fructose-1,6-bisphosphatase isozyme 2	Indirecta		
MAPK3_HUMAN	Q16644	MAP kinase-activated protein kinase 3	Indirecta		
XPA_HUMAN	P23025	DNA repair protein complementing XP-A cells	Indirecta	SI	
DYR1B_HUMAN	Q9Y463	Dual specificity tyrosine-phosphorylation-regulated kinase 1B	Indirecta		
GATA1_HUMAN	P15976	Erythroid transcription factor	Indirecta		
TFDP2_HUMAN	Q14188	Transcription factor Dp-2	Indirecta		SI
BAAT_HUMAN	Q14032	Bile acid-CoA:amino acid N-acyltransferase	Indirecta		
AP1AR_HUMAN	Q63HQ0	AP-1 complex-associated regulatory protein	Indirecta		
UHRF2_HUMAN	Q96PU4	E3 ubiquitin-protein ligase UHRF2	Indirecta		
NUCL_HUMAN	P19338	Nucleolin	Indirecta		
PABP2_HUMAN	Q86U42	Polyadenylate-binding protein 2	Indirecta		
MARF1_HUMAN	Q9Y4F3	Meiosis regulator and mRNA stability factor 1	Indirecta		
GSHR_HUMAN	P00390	Glutathione reductase, mitochondrial	Indirecta		
FOXO3_HUMAN	O43524	Forkhead box protein O3	Indirecta		
KDM5A_HUMAN	P29375	Lysine-specific demethylase 5A	Indirecta	SI	
F16P1_HUMAN	P09467	Fructose-1,6-bisphosphatase 1	Indirecta		
IRF3_HUMAN	Q14653	Interferon regulatory factor 3	Indirecta		
RASA1_HUMAN	P20936	Ras GTPase-activating protein 1	Indirecta		
DVL1_HUMAN	O14640	Segment polarity protein dishevelled homolog DVL-1	Indirecta		
TAF1_HUMAN	P21675	Transcription initiation factor TFIID subunit 1250	Indirecta	SI	
MDM4_HUMAN	O15151	Protein Mdm4	Indirecta		
LEF1_HUMAN	Q9UJU2	Lymphoid enhancer-binding factor 1	Indirecta		
KAT2B_HUMAN	Q92831	Histone acetyltransferase KAT2B	Indirecta		
RAF1_HUMAN	P04049	RAF proto-oncogene serine/threonine-protein kinase	Indirecta		

UniprotID	Accession	Nombre	Tipo de Interacción	Hit ProP-PD*	Proteómica**
HDAC2_HUMAN	Q92769	Histone deacetylase 2	Indirecta		SI
DYR1A_HUMAN	Q13627	Dual specificity tyrosine-phosphorylation-regulated kinase 1A	Indirecta		
GRB2_HUMAN	P62993	Growth factor receptor-bound protein 2	Indirecta		
SEPT4_HUMAN	O43236	Septin-4	Indirecta		
MDM2_HUMAN	Q00987	E3 ubiquitin-protein ligase Mdm2	Indirecta		
AGO2_HUMAN	Q9UKV8	Protein argonaute-2	Indirecta		
SHC1_HUMAN	P29353	SHC-transforming protein 1	Indirecta		
PSD10_HUMAN	O75832	26S proteasome non-ATPase regulatory subunit 10	Indirecta		SI

*Proteínas que contienen un péptido detectado en el ensayo ProP-PD

**Proteínas reportadas en ensayos de proteómica realizados con Rb [42,48].

Tabla S4. Interactores de la proteína p107 reportados en la base de datos de interactores, IntAct.

UniprotID	Accession	Nombre	Tipo de Interacción	Hit ProP-PD*	Proteómica**
CDK6_HUMAN	Q00534	Cyclin-dependent kinase 6	Directa		SI
CDK4_HUMAN	P11802	Cyclin-dependent kinase 4	Directa		
PP2AA_HUMAN	P67775	Serine/threonine-protein phosphatase 2A catalytic subunit alpha isoform	Directa-Indirecta		
HELZ_HUMAN	P42694	Probable helicase with zinc finger domain	Indirecta		
IRF3_HUMAN	Q14653	Interferon regulatory factor 3	Indirecta		
DYR1B_HUMAN	Q9Y463	Dual specificity tyrosine-phosphorylation-regulated kinase 1B	Indirecta		
TSN7_HUMAN	P41732	Tetraspanin-7	Indirecta		
CD320_HUMAN	Q9NPF0	CD320 antigen	Indirecta		
AOXA_HUMAN	Q06278	Aldehyde oxidase	Indirecta		
CD2B2_HUMAN	O95400	CD2 antigen cytoplasmic tail-binding protein 2	Indirecta		
DGKZ_HUMAN	Q13574	Diacylglycerol kinase zeta	Indirecta		
E2F1_HUMAN	Q01094	Transcription factor E2F1	Indirecta		SI
ARP3B_HUMAN	Q9P1U1	Actin-related protein 3B	Indirecta		
BEGIN_HUMAN	Q9BUH8	Brain-enriched guanylate kinase-associated protein	Indirecta		
FOXO3_HUMAN	O43524	Forkhead box protein O3	Indirecta		
DNPEP_HUMAN	Q9ULA0	Aspartyl aminopeptidase	Indirecta		
MARF1_HUMAN	Q9Y4F3	Meiosis regulator and mRNA stability factor 1	Indirecta		
TRIP6_HUMAN	Q15654	Thyroid receptor-interacting protein 6	Indirecta		
CTIP_HUMAN	Q99708	DNA endonuclease RBBP8	Indirecta		
LAMB2_HUMAN	P55268	Laminin subunit beta-2	Indirecta		
MOFA1_HUMAN	Q9Y605	MORF4 family-associated protein 1	Indirecta		
GOGA2_HUMAN	Q08379	Golgin subfamily A member 2	Indirecta		
E2F4_HUMAN	Q16254	Transcription factor E2F4	Indirecta	SI	SI
SMD3_HUMAN	P62318	Small nuclear ribonucleoprotein Sm D3	Indirecta		
RBL2_HUMAN	Q08999	Retinoblastoma-like protein 2	Indirecta		

UniprotID	Accession	Nombre	Tipo de Interacción	Hit ProP-PD*	Proteómica**
PLS1_HUMAN	O15162	Phospholipid scramblase 1	Indirecta		
NTH_HUMAN	P78549	Endonuclease III-like protein	Indirecta		
MK06_HUMAN	Q16659	Mitogen-activated protein kinase 6	Indirecta		
NUCB1_HUMAN	Q02818	Nucleobindin-1	Indirecta		
DYR1A_HUMAN	Q13627	Dual specificity tyrosine-phosphorylation-regulated kinase 1A	Indirecta		
EPHA2_HUMAN	P29317	Ephrin type-A receptor 2	Indirecta		SI
FINC_HUMAN	P02751	Fibronectin [Cleaved into: Anastellin	Indirecta		SI
NR4A1_HUMAN	P22736	Nuclear receptor subfamily 4immunitygroup A member 1	Indirecta		
DYL1_HUMAN	P63167	Dynein light chain 1, cytoplasmic	Indirecta		
1433Z_HUMAN	P63104	14-3-3 protein zeta/delta	Indirecta		
EPHB2_HUMAN	P29323	Ephrin type-B receptor 2	Indirecta		
NEK6_HUMAN	Q9HC98	Serine/threonine-protein kinase Nek6	Indirecta		

*Proteínas que contienen un péptido detectado en el ensayo ProP-PD

**Proteínas reportadas en ensayos de proteómica realizados con Rb [42,48].

Tabla S5. Variantes del SLiM LxCxE presentes en péptidos *hit* de Rb con mejores características en sus determinantes de unión

Uniprot ID	Proteína	Secuencia	Variante*
CPSF7_HUMAN	Cleavage and polyadenylation specificity factor subunit 7	GVDLIDIYADEEFNQD	1
ASB3_HUMAN	Ankyrin repeat and SOCS box protein 3	GADPDLYCNEDSWQLP	1
HFM1_HUMAN	Probable ATP-dependent DNA helicase HFM1	MLKSNDCLFSLENLFF	2A
HFM1_HUMAN	Probable ATP-dependent DNA helicase HFM1	LFSLENLFFEKPDEVE	2A
NCKX1_HUMAN	Sodium/potassium/calcium exchanger 1	SLFSREILNLTWWPLF	2A
ABCA8_HUMAN	ATP-binding cassette sub-family A member 8	SHLLFSSLFSEERMDV	2A
KDM5A_HUMAN	Lysine-specific demethylase 5A	EPNLFCDEEIPIKSEE	2A
KIF24_HUMAN	Kinesin-like protein KIF24	QSRETVLFSHEHMGSE	2A
HMBX1_HUMAN	Homeobox-containing protein 1	LHALETLDRLDQEHSD	2A
SUSD6_HUMAN	Sushi domain-containing protein 6	ALPSYEEAVYGSSGHC	2A
RBM33_HUMAN	RNA-binding protein 33	EEQLYTDEVLDIEINE	2A
TAF1_HUMAN	Transcription initiation factor TFIID subunit 1	SLITELTANEELTGTD	2B
SGSM1_HUMAN	Small G protein signaling modulator 1	SLESDLLANESMDEFM	2B
SGSM1_HUMAN	Small G protein signaling modulator 1	DLLANESMDEFMSITG	2B
EPHA1_HUMAN	Ephrin type-A receptor 1	PYVDLQAYEDPAQGAL	2C
SEPT7_HUMAN	Septin-7	SLFLTDLYSPEYPGPS	2C

*Variante 1: **[DE]**[IL]**[YFH]**[CAST].E.**{1,2}[WFILYVM]**o **[DE]**[IL].[CAST]**[YFH]**E.**{1,2}[WFILYVM];**

Variante 2A: [IL]**[YFH]**[CAST].E.**{1,2}[WFILYVM]**o [IL].[CAST]**[YFH]**E.**{1,2}[WFILYVM];**

Variante 2B: **[DE]**[IL].[CAST].E.**{1,2}[WFILYVM];**

Variante 2C: **[DE]**[IL]**[YFH]**[CAST].E o **[DE]**[IL].[CAST]**[YFH]**E.

Tabla S6. Variantes del SLiM LxCxE presentes en péptidos *hit* de p107 con mejores características en sus determinantes de unión

Uniprot ID	Proteína	Secuencia	Variante*
UBP10_HUMAN	Ubiquitin carboxyl-terminal hydrolase 10	GTATNGVELHTTESID	1
KIF15_HUMAN	Kinesin-like protein KIF15	STQMQELFSSERIDWT	1
KIF15_HUMAN	Kinesin-like protein KIF15	QELFSSERIDWTKQQE	1
PDLI1_HUMAN	PDZ and LIM domain protein 1	GLYSSENISNFNNALE	2A
NCKX1_HUMAN	Sodium/potassium/calcium exchanger 1	SLFSREILNLTWWPLF	2A
CAC1F_HUMAN	Voltage-dependent L-type calcium channel subunit alpha-1F	SSLYSDEESILSRFDE	2A
LDB3_HUMAN	LIM domain-binding protein 3	QYNNPIGLYSAETLRE	2A
ABCA8_HUMAN	ATP-binding cassette sub-family A member 8	SHLLFSSLFSEERMDV	2A
KDM5A_HUMAN	Lysine-specific demethylase 5A	EPNLFCDEEIPIKSEE	2A
LIN52_HUMAN	Protein lin-52 homolog	TDLEASLLSFEKLDRA	2A
KIF24_HUMAN	Kinesin-like protein KIF24	QSRETVLFSHEHMGSE	2A
UTP25_HUMAN	Digestive organ expansion factor homolog	SLFSLETNFLEEESGD	2A
HMBX1_HUMAN	Homeobox-containing protein 1	LHALETLDRLDQEHSD	2A
E2F4_HUMAN	Transcription factor E2F4	SELLEELMSSEVFAPL	2B
E2F4_HUMAN	Transcription factor E2F4	EELMSSEVFAPLLRLS	2B
SGSM1_HUMAN	Small G protein signaling modulator 1	DLLANESMDEFMSITG	2B
ZN436_HUMAN	Zinc finger protein 436	QWGDLTAEEWVSYPLQ	2B
PLAL1_HUMAN	Zinc finger protein PLAGL1	VCALELGSTEVLLDHL	2B
MNAR1_HUMAN	UPF0258 protein KIAA1024	KLTALDLQTQESLNPN	2B
SEPT7_HUMAN	Septin-7	SLFLTDLYSPEYPGPS	2C

*Variante 1: **[DE]**[IL]**[YFH]**[CAST].E.**{1,2}[WFILYVM]**o **[DE]**[IL].[CAST]**[YFH]**E.**{1,2}[WFILYVM];**

Variante 2A: [IL]**[YFH]**[CAST].E.**{1,2}[WFILYVM]**o [IL].[CAST]**[YFH]**E.**{1,2}[WFILYVM];**

Variante 2B: **[DE]**[IL].[CAST].E.**{1,2}[WFILYVM];**

Variante 2C: **[DE]**[IL]**[YFH]**[CAST].E o **[DE]**[IL].[CAST]**[YFH]**E.

Tabla S7. Variantes del SLiM E2F presentes en péptidos *hit* de Rb con mejores características en sus determinantes de unión

Uniprot ID	Proteína	Secuencia	Variante*
ARB2P_HUMAN	Putative protein FAM172B	TAYIWDYFISKTEGKD	1
HAX1_HUMAN	HCLS1-associated protein X-1	PALDDAFSILDLFLGR	1
E2F3_HUMAN	Transcription factor E2F3	DYLLSLGEEEGISDLF	1
AP180_HUMAN	Clathrin coat assembly protein AP180	AKVDSSGVIDLFGDAF	1
AP180_HUMAN	Clathrin coat assembly protein AP180	SSGVIDLFGDAFGSSA	1
SYDE2_HUMAN	Rho GTPase-activating protein SYDE2	VLSVPPDQRITLTDLF	1
CC190_HUMAN	Coiled-coil domain-containing protein 190	SERLLSIGEIFGHGES	1
DYH10_HUMAN	Dynein heavy chain 10, axonemal	DIILLSEMFSDNFGQL	1
CPSF7_HUMAN	Cleavage and polyadenylation specificity factor subunit 7	GVDLIDIYADEEFNQD	1
EM55_HUMAN	55 kDa erythrocyte membrane protein	GSMHTALSDLYLEHLL	1
E2F1_HUMAN	Transcription factor E2F1	DYHFGLEEGEGIRDLF	1
CDR2_HUMAN	Cerebellar degeneration-related protein 2	KEPSQSLLEEMFLTVP	1
E2F2_HUMAN	Transcription factor E2F2	DYLWGLEAGEGISDLF	1
E2F2_HUMAN	Transcription factor E2F2	GLEAGEGISDLFDSYD	1
E2F2_HUMAN	Transcription factor E2F2	GEGISDLFDSYDLGDL	1
EPN4_HUMAN	Clathrin interactor 1	LVDLFDGTSQSTGGSA	1
EMC2_HUMAN	ER membrane protein complex subunit 2	VSELYDVTWEEMRDKM	1
SEPT2_HUMAN	Septin-2	SLFLTDLYPERVIPGA	1
E2F5_HUMAN	Transcription factor E2F5	YNFNLDDNEGVCDLFD	1
SEPT7_HUMAN	Septin-7	SLFLTDLYSPEYPGPS	1
RHG17_HUMAN	Rho GTPase-activating protein 17	DWFFPEEVEFNVSEAF	1
NRK_HUMAN	Nik-related protein kinase	NALSEIFRNDWLTPAP	1

HUWE1_HUMAN	E3 ubiquitin-protein ligase HUWE1	LLDDFFHDQSTATSQA	1
CC190_HUMAN	Coiled-coil domain-containing protein 190	LSIGEIFGHGESSSSR	1
ERO1B_HUMAN	ERO1-like protein beta	FERSIVDLYTGNAEED	1
DYH10_HUMAN	Dynein heavy chain 10, axonemal	QGWEDIILLSEMFSDN	1
NALCN_HUMAN	Sodium leak channel non-selective protein	TPTAVIRDFGGVMDIF	1
AGGF1_HUMAN	Angiogenic factor with G patch and FHA domains 1	EVQTENHAPWSISDYF	1
PRUN2_HUMAN	Protein prune homolog 2	SGIMELYGSDIEPQPS	1
TPC12_HUMAN	Trafficking protein particle complex subunit 12	ASDFFDSFTTSAFISV	1
KCNB2_HUMAN	Potassium voltage-gated channel subfamily B member 2	LDYWGIDEIYLESCCQ	1
ZN317_HUMAN	Zinc finger protein 317	HLFEVFGMDPHLTQPM	1
GON7_HUMAN	EKC/KEOPS complex subunit GON7	VTELFDPLVQGEVQHR	1
EPN3_HUMAN	Epsin-3	SQSSILDLADIFVPAL	1
PKHG2_HUMAN	Pleckstrin homology domain-containing family G member 2	ISDVFEMPCLPAIPSV	1
INCE_HUMAN	Inner centromere protein	YHPPNLLELFGTILPL	1
VPS54_HUMAN	Vacuolar protein sorting-associated protein 54	EVAYLIHEGMFISDAF	1
VPS54_HUMAN	Vacuolar protein sorting-associated protein 54	LIHEGMFISDAFGEGE	1
VPS54_HUMAN	Vacuolar protein sorting-associated protein 54	GMFISDAFGEGELTPI	1
STAR9_HUMAN	StAR-related lipid transfer protein 9	DSLNAKLEGVSDFFST	1
STAR9_HUMAN	StAR-related lipid transfer protein 9	GVSDFFSTSEKEASYD	1
I20RA_HUMAN	Interleukin-20 receptor subunit alpha	LGYASHLMEIFCDSEE	1
MYT1L_HUMAN	Myelin transcription factor 1-like protein	YVTTLTEMYTNQDRYQ	1
SCAF8_HUMAN	SR-related and CTD-associated factor 8	VFDYFEGATSQRKGDN	1
GRIP1_HUMAN	Glutamate receptor-interacting protein 1	LSDMYPSTVPSVDSAV	1

* Variante 1: `[IVLMA].[DE][IVLMAFYW][FY]`

Tabla S8. Variantes del SLiM E2F presentes en péptidos *hit* de p107 con mejores características en sus determinantes de unión

Uniprot ID	Proteína	Secuencia	Variante*
E2F3_HUMAN	Transcription factor E2F3	SDLFDAYDLEKLPLVE	1
SEPT7_HUMAN	Septin-7	SLFLTDLYSPEYPGPS	1
E2F2_HUMAN	Transcription factor E2F2	DYLWGLEAGEGISDLF	1
KIF15_HUMAN	Kinesin-like protein KIF15	STQMQELFSSERIDWT	1
MYT1L_HUMAN	Myelin transcription factor 1-like protein	YVTTLTEMYTNQDRYQ	1

* Variante 1: [IVLMA]**.[DE]**[IVLMAFYW]**[FY]**

Tabla S9. Matriz FoldX del complejo E7- Rb (PDB: 1GUX [5]).

1GUX																				
Pos	Res	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
1	**D**	0,176	0,339	0,206	0,000	0,148	0,236	0,043	0,193	0,353	0,271	0,294	0,218	-0,031	0,562	0,246	0,230	0,222	0,239	0,331	0,229
2	**L**	3,875	3,248	2,952	4,536	3,126	2,534	4,152	5,194	4,303	2,406	0,000	3,633	0,392	0,136	3,197	4,984	3,811	6,093	5,109	2,511
3	**Y**	0,914	0,793	1,028	1,054	0,814	0,771	0,755	1,200	0,554	0,126	-0,408	0,276	-0,435	-0,481	1,096	1,192	0,943	-0,281	0,000	0,419
4	**C**	0,110	7,258	1,109	0,692	0,000	3,988	2,810	1,699	25,975	6,702	7,664	6,427	2,102	30,966	1,773	1,583	1,289	32,170	32,452	0,516
5	**Y**	0,013	0,444	0,063	-0,186	0,025	0,187	-0,171	0,004	0,048	0,184	0,011	0,066	-0,012	-0,017	0,037	-0,019	0,139	0,193	0,000	0,195
6	**E**	3,394	4,351	3,254	3,197	2,851	3,198	0,000	4,067	4,638	2,141	2,055	4,106	3,262	4,518	5,283	3,546	3,552	4,485	4,562	2,185
7	**Q**	0,114	0,191	0,261	-0,037	0,078	0,000	-0,031	-0,146	0,179	0,040	0,067	0,239	0,035	-0,077	0,108	0,107	0,045	-0,046	-0,016	0,032
8	**L**	3,026	5,242	2,411	1,934	2,010	2,911	3,117	3,620	12,105	1,642	0,000	3,488	-0,544	2,037	3,226	3,292	2,239	9,201	4,970	2,144
9	**N**	-0,178	0,184	0,000	-0,293	-0,083	0,005	-0,178	-0,563	0,117	0,130	0,186	0,653	-0,017	0,059	3,578	-0,238	-0,155	-0,006	0,043	0,134

* Posiciones del residuo de la matriz. ** Residuo identificado en el complejo.

Tabla S10. Matriz FoldX del complejo LIN52-p107 (PDB: 4YOS [22]**).**

4YOS																				
Pos	Res	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
1	**S**	-0,008	-0,488	-0,139	-0,128	-0,118	-0,281	-0,374	-0,033	-0,294	-0,325	-0,651	-0,639	-1,160	-0,972	0,014	0,000	0,020	-1,256	-0,973	-0,074
2	**L**	2,221	1,930	2,622	2,189	2,028	2,224	1,384	2,339	1,585	0,995	0,000	1,927	-0,241	-0,673	1,734	2,882	2,462	1,876	0,375	1,417
3	**L**	1,005	0,206	1,413	1,335	0,498	-0,116	0,265	1,639	0,595	0,356	0,000	-0,032	0,172	1,696	0,998	1,386	1,126	1,745	1,506	0,583
4	**S**	-0,638	-0,250	-1,039	-0,867	-0,621	-1,035	-1,331	-0,281	-0,820	1,052	-0,929	-0,628	-1,577	-1,391	0,145	0,000	-2,193	-1,152	-1,027	0,545
5	**F**	0,513	0,665	0,659	-0,738	0,379	0,380	-0,300	0,561	0,490	0,218	0,133	0,168	-0,836	0,000	0,435	0,636	0,617	-0,251	-0,508	0,456
6	**E**	0,068	0,155	0,045	-1,063	-0,592	-0,049	0,000	0,277	0,004	-0,678	-1,394	0,023	-1,854	0,251	0,433	0,215	-0,572	-0,382	0,333	-0,310
7	**K**	0,111	0,067	0,131	0,091	0,079	0,096	0,235	0,160	0,088	0,067	0,073	0,000	-0,015	-0,032	0,283	0,122	0,074	0,048	-0,035	0,068
8	**L**	1,201	1,928	0,671	0,813	1,088	0,851	1,340	1,634	0,986	0,549	0,000	1,262	-0,796	-2,427	0,738	2,099	2,253	-2,460	-0,553	0,912
9	**D**	0,237	0,346	0,250	0,000	0,250	0,198	-0,120	0,303	0,253	0,279	0,178	0,342	0,154	0,151	0,196	0,266	0,250	-0,762	0,142	0,269

* Posiciones del residuo de la matriz. ** Residuo identificado en el complejo.

Tabla S11. Matriz FoldX del complejo E1A-Rb (PDB: 2R7G [33]**).**

2R7G																				
Pos*	**Res****	**ALA**	**ARG**	**ASN**	**ASP**	**CYS**	**GLN**	**GLU**	**GLY**	**HIS**	**ILE**	**LEU**	**LYS**	**MET**	**PHE**	**PRO**	**SER**	**THR**	**TRP**	**TYR**	**VAL**
1	**P**	0,018	1,288	0,235	1,245	0,199	0,574	1,352	0,476	0,369	-0,868	-0,209	0,787	-0,925	-1,040	0,000	-0,285	0,196	-0,863	-0,429	-0,529
2	**P**	-0,424	-1,532	-0,402	0,614	-0,294	-0,667	-0,626	0,179	0,467	-1,001	-1,067	-0,774	-1,312	-2,322	0,000	-0,469	-0,469	-1,034	-2,303	-1,074
3	**T**	0,229	0,232	-0,027	0,885	-0,128	0,491	0,127	0,531	0,147	-0,614	-0,859	0,351	-0,024	-1,326	0,300	0,699	0,000	0,518	-0,385	-0,853
4	**L**	3,305	0,219	3,854	4,715	2,999	2,812	3,946	4,065	0,640	0,555	0,000	1,266	-0,472	0,997	5,990	4,061	3,546	2,558	6,482	1,866
5	**H**	1,741	1,261	1,038	2,659	1,512	1,805	2,502	1,936	0,000	1,365	0,578	1,300	0,024	-0,521	3,708	2,117	1,337	1,086	2,024	1,955
6	**E**	0,049	-0,355	0,052	0,248	-0,281	-0,046	0,000	0,489	-0,178	-0,121	-1,204	-0,801	-1,450	-1,065	1,746	0,379	0,093	-1,390	-0,687	-0,202
7	**L**	1,516	-0,637	1,135	2,455	1,420	0,938	2,478	2,172	-0,016	0,637	0,000	-0,044	-0,503	-1,412	5,340	1,926	1,938	-2,646	-0,850	1,439
8	**Y**	3,117	2,925	1,603	3,630	2,799	2,239	3,595	3,435	1,350	1,497	1,716	2,351	1,275	-1,390	7,276	2,709	2,586	0,272	0,000	2,122
9	**D**	4,368	4,905	2,861	0,000	3,370	3,766	3,369	4,134	4,786	3,835	3,529	4,628	3,199	4,389	5,772	4,763	4,256	4,066	4,479	4,310
10	**L**	1,297	1,198	1,479	1,670	1,231	2,352	2,227	1,363	0,833	0,147	0,000	0,905	-0,331	-0,976	5,876	1,930	2,072	0,956	-0,482	1,007

* Posiciones del residuo de la matriz. ** Residuo identificado en el complejo.

Tabla S12. Matriz FoldX del complejo E2F2-Rb (PDB: 1N4M [38]**).**

	1N4M																			
Pos	Res	**ALA**	**ARG**	**ASN**	**ASP**	**CYS**	**GLN**	**GLU**	**GLY**	**HIS**	**ILE**	**LEU**	**LYS**	**MET**	**PHE**	**PRO**	**SER**	**THR**	**TRP**	**TYR**	**VAL**
1	**G**	1,370	0,412	0,556	0,700	0,230	0,035	0,576	0,000	1,070	1,639	0,446	-0,416	0,405	0,820	2,317	0,295	1,632	0,203	0,534	1,882
2	**E**	-0,399	-1,163	-0,311	-0,879	-0,586	-0,451	0,000	-0,405	-0,470	-0,287	-1,704	-0,874	-1,918	-0,892	2,426	-0,367	-0,742	-2,096	-0,735	-0,126
3	**G**	2,732	2,833	3,738	3,923	2,369	2,472	3,443	0,000	4,805	4,158	2,668	1,986	0,970	5,561	4,129	2,803	3,103	6,864	5,825	5,440
4	**I**	2,648	3,909	3,160	3,612	2,497	1,754	3,841	3,695	4,645	0,000	2,720	2,861	-0,038	6,047	2,162	2,833	2,529	6,537	6,486	0,387
5	**S**	-0,114	-0,104	0,273	0,694	-0,432	0,117	0,337	0,490	1,364	-0,461	-1,641	0,465	-1,638	1,396	1,238	0,000	-1,179	1,495	1,297	-0,034
6	**D**	1,161	0,981	-0,571	0,000	0,897	0,972	0,478	1,527	1,129	1,630	0,907	0,687	0,895	0,208	5,063	1,314	2,334	0,954	0,373	1,854
7	**L**	1,717	-0,094	0,414	3,700	1,014	1,228	1,958	2,264	0,061	1,292	0,000	0,125	-0,998	-0,925	3,974	2,037	0,903	-0,415	-0,627	0,984
8	**F**	3,511	3,231	2,400	3,924	3,163	3,479	4,696	4,116	2,046	2,531	1,466	1,707	0,881	0,000	4,926	4,262	3,784	2,570	0,159	2,547
9	**D**	-0,284	-0,496	-0,057	0,000	-0,342	-0,173	-0,002	-0,718	0,024	0,367	0,133	-0,652	0,022	-0,130	-0,033	-1,056	-0,748	0,197	-0,044	0,324

* Posiciones del residuo de la matriz. ** Residuo identificado en el complejo.

Tabla S13. Interactores conocidos (TP) reportados en ELM de la proteína *pocket* Rb escaneados con la matriz FoldX 1GUX_8.

Uniprot ID	Proteína	Instancia	Secuencia	Sub-secuencia*	FoldX**
B8ZX42_9POLY	Large T antigen	True Positive	TWEDLFCDESLSSPEPPS	LFCDESLS	-0.8
CCD11_ARATH	Cyclin-D1-1	True Positive	NDMDLFCGEDSGVFSGES	LFCGEDSG	2.21
CCD21_ARATH	Cyclin-D2-1	True Positive	MAENLACGETSESWIIDN	LACGETSE	4.08
CCD31_ARATH	Cyclin-D3-1	True Positive	LLDALYCEEEKWDDEGEE	LYCEEEKW	3.28
CLINK_FBNY1	Cell cycle link protein	True Positive	DMDDLSCRELLPPEEDDD	LSCRELLP	5.28
E1A_ADE05	Early E1A protein	True Positive	EVIDLTCHEAGFPPSDDE	LTCHEAGF	4.78
VE7_HPV11	Protein E7	True Positive	DPVGLHCYEQLEDSSEDE	LHCYEQLE	0.38
E7BRQ8_9PAPI	Protein E7	True Positive	LPANLLSTESLSPDDELE	LLSTESLS	1.18
VE7_HPV16	Protein E7	True Positive	ETTDLYCYEQLNDSSEEE	LYCYEQLN	0
VE7_HPV18	Protein E7	True Positive	IPVDLLCHEQLSDSEEEN	LLCHEQLS	-0.6
VE7_HPV20	Protein E7	True Positive	QPVDLFCEEELPNEQQER	LFCEEELP	2.89
VE7_HPV22	Protein E7	True Positive	LPIDLHCHEELPELPEEL	LHCHEELP	4.15
VE7_HPV31	Protein E7	True Positive	EATDLHCYEQLPDSSDEE	LHCYEQLP	4.13
VE7_HPV38	Protein E7	True Positive	QPIDLHCHEELPDLPEDI	LHCHEELP	4.15
VE7_HPV04	Protein E7	True Positive	LPANLLSEEVLQSSDDEY	LLSEEVLQ	01.04
VE7_HPV50	Protein E7	True Positive	LPVNLLSDESIETDDIAE	LLSDESIE	2.56

Uniprot ID	Proteína	Instancia	Secuencia	Sub-secuencia*	FoldX**
VE7_HPV52	Protein E7	True Positive	ETTDLHCYEQLGDSSDEE	LHCYEQLG	-0.01
UL97_HCMVA	Serine/threonine protein kinase UL97	True Positive	GYHGLRCRETSAMWSFEY	LRCRETSA	4.4
HBP1_RAT	HMG box-containing protein 1	True Positive	SLELLQCNENVPSSPGYN	LQCNENVP	6.82
LT_SV40	Large T antigen	True Positive	NEENLFCSEEMPSSDDEA	LFCSEEMP	2.5
Q63928_9MURI	Brg1 protein	True Positive	EVERLTCEEEEEKMFGRG	LTCEEEEE	3.68
Q77J89_WSSVS	Wsv069	True Positive	GHSDLTCDEISGFFVAAA	LTCDEISG	3.53
Q77J94_WSSVS	Wsv056	True Positive	GHSDLSCDEISEFLVQAA	LSCDEISE	4.16
MC007_MCV1	Protein MC007	True Positive	VEVDLYCHENLRYESDVS	LYCHENLR	0.49
MC007_MCV1	Protein MC007	True Positive	LLDLLACSEDASGFSPPE	LACSEDAS	3.65
Q9WKM8_BBTV	Cell cycle link protein	True Positive	VYQDLYCDEVLSSSSTEE	LYCDEVLS	-0.39
REPA_BEYDV	Replication-associated protein A	True Positive	TESDLRCHEDLHNWRETH	LRCHEDLH	0.92
REPA_MSVS	Replication-associated protein A	True Positive	SSPDLLCNESINDWLQPN	LLCNESIN	1.4
REPA_WDVS	Replication-associated protein A	True Positive	PTESLICHETIESWKNEH	LICHETIE	1.68
ARI4A_HUMAN	AT-rich interactive domain-containing protein 4A	True Negative	GPETAVAHAVDLDDLDEK	AVAHAVDL	10
B8ZX42_9POLY	Large T antigen	True Negative	TWEDAFADASLSSPEPPS	AFADASLS	6.58
CCD11_ARATH	Cyclin-D1-1	True Negative	NDMDAFAGADSGVFSGES	AFAGADSG	9.59
CCD21_ARATH	Cyclin-D2-1	True Negative	MAENAAAGATSESWIIDN	AAAGATSE	11.46
CCD31_ARATH	Cyclin-D3-1	True Negative	LLDAAYAEAEKWDDEGEE	AYAEAEKW	10.66

Uniprot ID	Proteína	Instancia	Secuencia	Sub-secuencia*	FoldX**
CCND1_HUMAN	G1/S-specific cyclin-D1	True Negative	MEHQALACAVETIRRAYP	ALACAVET	9.99
CCND2_HUMAN	G1/S-specific cyclin-D2	True Negative	AAMEALAHAVDPVRRAVR	ALAHAVDP	12.56
CCND3_HUMAN	G1/S-specific cyclin-D3	True Negative	AAMEALACAGTRHAPRAG	ALACAGTR	9.27
CLINK_FBNY1	Cell cycle link protein	True Negative	DMDDASARALLPPEEDDD	ASARALLP	12.66
E1A_ADE05	Early E1A protein	True Negative	EVIDATAHAAGFPPSDDE	ATAHAAGF	12.16
VE7_HPV11	Protein E7	True Negative	DPVGAHAYAQLEDSSEDE	AHAYAQLE	7.76
E7BRQ8_9PAPI	Protein E7	True Negative	LPANALATASLSPDDELE	ALATASLS	6.98
VE7_HPV16	Protein E7	True Negative	ETTDAYAYAQLNDSSEEE	AYAYAQLN	7.38
VE7_HPV18	Protein E7	True Negative	IPVDALAHAQLSDSEEEN	ALAHAQLS	6.78
VE7_HPV20	Protein E7	True Negative	QPVDAFAEAELPNEQQER	AFAEAELP	10.27
VE7_HPV22	Protein E7	True Negative	LPIDAHAHAELPELPEEL	AHAHAELP	11.53
VE7_HPV31	Protein E7	True Negative	EATDAHAYAQLPDSSDEE	AHAYAQLP	11.51
VE7_HPV38	Protein E7	True Negative	QPIDAHAHAELPDLPEDI	AHAHAELP	11.53
VE7_HPV04	Protein E7	True Negative	LPANALAEAVLQSSDDEY	ALAEAVLQ	6.84
VE7_HPV50	Protein E7	True Negative	LPVNALADASIETDDIAE	ALADASIE	8.36
VE7_HPV52	Protein E7	True Negative	ETTDAHAYAQLGDSSDEE	AHAYAQLG	7.37
EID1_HUMAN	EP300-interacting inhibitor of differentiation 1	True Negative	LTEEAGADAIIDREAAAA	AGADAIID	9.78
UL97_HCMVA	Serine/threonine protein kinase UL97	True Negative	GYHGARARATSAMWSFEY	ARARATSA	11.77

Uniprot ID	Proteína	Instancia	Secuencia	Sub-secuencia*	FoldX**
HBP1_RAT	HMG box-containing protein 1	True Negative	SLELAQANANVPSSPGYN	AQANANVP	14.2
HDAC1_HUMAN	Histone deacetylase 1	True Negative	SDKRAAAEAEFSDSEEEG	AAAEAEFS	9.89
HDAC2_HUMAN	Histone deacetylase 2	True Negative	SDKRAAADAEFSDSEDEG	AAADAEFS	9.88
KDM5A_HUMAN	Lysine-specific demethylase 5A	True Negative	LEPNAFADAEIPIKSEEV	AFADAEIP	11.9
LT_SV40	Large T antigen	True Negative	NEENAFASAEMPSSDDEA	AFASAEMP	9.88
NDC80_HUMAN	Kinetochore protein NDC80 homolog	True Negative	LDYTAKAYASFMSGADSF	AKAYASFM	9.78
PPR26_HUMAN	Protein phosphatase 1 regulatory subunit 26	True Negative	TSAEAMAAAAILDISKTI	AMAAAAIL	8.9
PRDM2_HUMAN	PR domain zinc finger protein 2	True Negative	KEPEARADAKPEDLLEEP	ARADAKPE	11.27
Q63928_9MURI	Brg1 protein	True Negative	EVERATAEAEEEKMFGRG	ATAEAEEE	11.06
Q77J89_WSSVS	Wsv069	True Negative	GHSDATADAISGFFVAAA	ATADAISG	10.9
Q77J94_WSSVS	Wsv056	True Negative	GHSDASADAISEFLVQAA	ASADAISE	11.54
MC007_MCV1	Protein MC007	True Negative	VEVDAYAHANLRYESDVS	AYAHANLR	7.87
MC007_MCV1	Protein MC007	True Negative	LLDLAAASADASGFSPPE	AAASADAS	11.03
Q9WKM8_BBTV	Cell cycle link protein	True Negative	VYQDAYADAVLSSSSTEE	AYADAVLS	6.99
REPA_BEYDV	Replication-associated protein A	True Negative	TESDARAHADLHNWRETH	ARAHADLH	8.3
REPA_MSVS	Replication-associated protein A	True Negative	SSPDALANASINDWLQPN	ALANASIN	8.78
REPA_WDVS	Replication-associated protein A	True Negative	PTESAIAHATIESWKNEH	AIAHATIE	09.06
SMCA2_HUMAN	Probable global transcription activator SNF2L2	True Negative	EVERATAEAEEEKIFGRG	ATAEAEEE	11.06

Uniprot ID	Proteína	Instancia	Secuencia	Sub-secuencia*	FoldX**
SMCA4_HUMAN	Transcription activator BRG1	True Negative	EVERATAEAEEEKMFGRG	ATAEAEEE	11.06

*Sub-secuencia en la que se detecta la expresión regular `[IL].[CAST].E` en instancias True Positive; o misma sub-secuencia con mutaciones de A en instancias TN

**Valor de Foldx obtenido para la subsecuencia al ser escaneada con 1GUX_8

Tabla S14. Interactores conocidos (TP) reportados en ELM de la proteína *pocket* p107 escaneados con la matriz FoldX 1GUX_8.

Uniprot ID	Proteína	Instancia	Secuencia	Sub-secuencia*	FoldX**
B8ZX42_9POLY	Large T antigen	True Positive	TWEDLFCDESLSSPEPPS	LFCDESLS	-0.8
CCD11_ARATH	Cyclin-D1-1	True Positive	NDMDLFCGEDSGVFSGES	LFCGEDSG	2.21
CCD21_ARATH	Cyclin-D2-1	True Positive	MAENLACGETSESWIIDN	LACGETSE	04.08
CCD31_ARATH	Cyclin-D3-1	True Positive	LLDALYCEEEKWDDEGEE	LYCEEEKW	3.28
CLINK_FBNY1	Cell cycle link protein	True Positive	DMDDLSCRELLPPEEDDD	LSCRELLP	5.28
E1A_ADE05	Early E1A protein	True Positive	EVIDLTCHEAGFPPSDDE	LTCHEAGF	4.78
VE7_HPV11	Protein E7	True Positive	DPVGLHCYEQLEDSSEDE	LHCYEQLE	0.38
E7BRQ8_9PAPI	Protein E7	True Positive	LPANLLSTESLSPDDELE	LLSTESLS	1.18
VE7_HPV16	Protein E7	True Positive	ETTDLYCYEQLNDSSEEE	LYCYEQLN	0
VE7_HPV20	Protein E7	True Positive	QPVDLFCEEELPNEQQER	LFCEEELP	2.89
VE7_HPV22	Protein E7	True Positive	LPIDLHCHEELPELPEEL	LHCHEELP	4.15
VE7_HPV31	Protein E7	True Positive	EATDLHCYEQLPDSSDEE	LHCYEQLP	4.13
VE7_HPV04	Protein E7	True Positive	LPANLLSEEVLQSSDDEY	LLSEEVLQ	01.04
VE7_HPV50	Protein E7	True Positive	LPVNLLSDESIETDDIAE	LLSDESIE	2.56
VE7_HPV52	Protein E7	True Positive	ETTDLHCYEQLGDSSDEE	LHCYEQLG	-0.01
UL97_HCMVA	Serine/threonine protein kinase UL97	True Positive	GYHGLRCRETSAMWSFEY	LRCRETSA	4.4
LT_SV40	Large T antigen	True Positive	NEENLFCSEEMPSSDDEA	LFCSEEMP	2.5
PRDM2_HUMAN	PR domain zinc finger protein 2	True Positive	KEPEIRCDEKPEDLLEEP	IRCDEKPE	6.3
Q63928_9MURI	Brg1 protein	True Positive	EVERLTCEEEEEKMFGRG	LTCEEEEE	3.68
Q77J94_WSSVS	Wsv056	True Positive	GHSDLSCDEISEFLVQAA	LSCDEISE	4.16
MC007_MCV1	Protein MC007	True Positive	LLDLLACSEDASGFSPPE	LACSEDAS	3.65

Uniprot ID	Proteína	Instancia	Secuencia	Sub-secuencia*	FoldX**
Q9WKM8_BBTV	Cell cycle link protein	True Positive	VYQDLYCDEVLSSSSTEE	LYCDEVLS	-0.39
REPA_MSVS	Replication-associated protein A	True Positive	SSPDLLCNESINDWLQPN	LLCNESIN	1.4
ARI4A_HUMAN	AT-rich interactive domain-containing protein 4A	True Negative	GPETAVAHAVDLDDLDEK	AVAHAVDL	10
B8ZX42_9POLY	Large T antigen	True Negative	TWEDAFADASLSSPEPPS	AFADASLS	6.58
CCD11_ARATH	Cyclin-D1-1	True Negative	NDMDAFAGADSGVFSGES	AFAGADSG	9.59
CCD21_ARATH	Cyclin-D2-1	True Negative	MAENAAAGATSESWIIDN	AAAGATSE	11.46
CCD31_ARATH	Cyclin-D3-1	True Negative	LLDAAYAEAEKWDDEGEE	AYAEAEKW	10.66
CCND1_HUMAN	G1/S-specific cyclin-D1	True Negative	MEHQALACAVETIRRAYP	ALACAVET	9.99
CCND2_HUMAN	G1/S-specific cyclin-D2	True Negative	AAMEALAHAVDPVRRAVR	ALAHAVDP	12.56
CCND3_HUMAN	G1/S-specific cyclin-D3	True Negative	AAMEALACAGTRHAPRAG	ALACAGTR	9.27
CLINK_FBNY1	Cell cycle link protein	True Negative	DMDDASARALLPPEEDDD	ASARALLP	12.66
E1A_ADE05	Early E1A protein	True Negative	EVIDATAHAAGFPPSDDE	ATAHAAGF	12.16
VE7_HPV11	Protein E7	True Negative	DPVGAHAYAQLEDSSEDE	AHAYAQLE	7.76
E7BRQ8_9PAPI	Protein E7	True Negative	LPANALATASLSPDDELE	ALATASLS	6.98
VE7_HPV16	Protein E7	True Negative	ETTDAYAYAQLNDSSEEE	AYAYAQLN	7.38
VE7_HPV20	Protein E7	True Negative	QPVDAFAEAELPNEQQER	AFAEAELP	10.27
VE7_HPV22	Protein E7	True Negative	LPIDAHAHAELPELPEEL	AHAHAELP	11.53
VE7_HPV31	Protein E7	True Negative	EATDAHAYAQLPDSSDEE	AHAYAQLP	11.51
VE7_HPV04	Protein E7	True Negative	LPANALAEAVLQSSDDEY	ALAEAVLQ	6.84
VE7_HPV50	Protein E7	True Negative	LPVNALADASIETDDIAE	ALADASIE	8.36
VE7_HPV52	Protein E7	True Negative	ETTDAHAYAQLGDSSDEE	AHAYAQLG	7.37
EID1_HUMAN	EP300-interacting inhibitor of differentiation 1	True Negative	LTEEAGADAIIDREAAAA	AGADAIID	9.78

Uniprot ID	Proteína	Instancia	Secuencia	Sub-secuencia*	FoldX**
UL97_HCMVA	Serine/threonine protein kinase UL97	True Negative	GYHGARARATSAMWSFEY	ARARATSA	11.77
HDAC1_HUMAN	Histone deacetylase 1	True Negative	SDKRAAAEAEFSDSEEEG	AAAEAEFS	9.89
KDM5A_HUMAN	Lysine-specific demethylase 5A	True Negative	LEPNAFADAEIPIKSEEV	AFADAEIP	11.9
LIN52_HUMAN	Protein lin-52 homolog	True Negative	LEASALAFAKLDRASPDL	ALAFAKLD	6.9
LT_SV40	Large T antigen	True Negative	NEENAFASAEMPSSDDEA	AFASAEMP	9.88
PRDM2_HUMAN	PR domain zinc finger protein 2	True Negative	KEPEARADAKPEDLLEEP	ARADAKPE	11.27
Q63928_9MURI	Brg1 protein	True Negative	EVERATAEAEEEKMFGRG	ATAEAEEE	11.06
Q77J94_WSSVS	Wsv056	True Negative	GHSDASADAISEFLVQAA	ASADAISE	11.54
MC007_MCV1	Protein MC007	True Negative	LLDLAAASADASGFSPPE	AAASADAS	11.03
Q9WKM8_BBTV	Cell cycle link protein	True Negative	VYQDAYADAVLSSSSTEE	AYADAVLS	6.99
REPA_MSVS	Replication-associated protein A	True Negative	SSPDALANASINDWLQPN	ALANASIN	8.78
SMCA2_HUMAN	Probable global transcription activator SNF2L2	True Negative	EVERATAEAEEEKIFGRG	ATAEAEEE	11.06
SMCA4_HUMAN	Transcription activator BRG1	True Negative	EVERATAEAEEEKMFGRG	ATAEAEEE	11.06

*Sub-secuencia en la que se detecta la expresión regular `[IL].[CAST].E` en instancias True Positive; o misma sub-secuencia con mutaciones de A en instancias TN

**Valor de Foldx obtenido para la subsecuencia al ser escaneada con 1GUX_8

Tabla S15. Interactores conocidos (TP) reportados en ELM de la proteína *pocket* Rb escaneados con la matriz FoldX 1N4M_5.

Uniprot ID	Proteína	Instancia	Secuencia	Sub-secuencia*	FoldX**
E1A_ADE05	Early E1A protein	True Positive	SHFEPPTLHELYDLDVTA	LHELY	4.72
RB_HUMAN	Retinoblastoma-associated protein	True Positive	PLQNNHTAADMYLSPVRS	AADMY	1.7
E1A_ADE05	Early E1A protein	True Negative	SHFEPPTAHAAADLDVTA	AHAAA	10.4
E2F1_HUMAN	Transcription factor E2F1	True Negative	LDAAAGAEEGEGARAAAD	ARAAA	8.93
E2F2_HUMAN	Transcription factor E2F2	True Negative	DDAAAGAEAGEGASAAAD	ASAAA	9.04
E2F3_HUMAN	Transcription factor E2F3	True Negative	EDAAASAGEEEGASAAAD	ASAAA	9.04
E2F4_HUMAN	Transcription factor E2F4	True Negative	HDAAANADESEGACAAAD	ACAAA	8.61
E2F5_HUMAN	Transcription factor E2F5	True Negative	DDAAANADDNEGACAAAD	DDNEG	9.81
RB_HUMAN	Retinoblastoma-associated protein	True Negative	PLQNNHTAAAAALSPVRS	AAAAA	8.92

**Sub-secuencia con mutaciones de A en instancias TN
**Valor de Foldx obtenido para la subsecuencia al ser escaneada con 1N4M_5

Tabla S16. Interactores conocidos (TP) reportados en ELM de la proteína *pocket* p107 escaneados con la matriz FoldX 1N4M_5.

Uniprot ID	Proteína	Instancia	Secuencia	Sub-secuencia*	FoldX**
E2F1_HUMAN	Transcription factor E2F1	True Negative	LDAAAGAEEGEGARAAAD	ARAAA	8.93
E2F2_HUMAN	Transcription factor E2F2	True Negative	DDAAAGAEAGEGASAAAD	ASAAA	9.04
E2F3_HUMAN	Transcription factor E2F3	True Negative	EDAAASAGEEEGASAAAD	ASAAA	9.04
E2F4_HUMAN	Transcription factor E2F4	True Negative	HDAAANADESEGACAAAD	ACAAA	8.61

*Sub-secuencia con mutaciones de A en instancias TN

**Valor de Foldx obtenido para la subsecuencia al ser escaneada con 1N4M_5

Tabla S17. Hits ProP-PD de la proteína Rb con SLiM LxCxE testeados experimentalmente escaneados con 1GUX_8.

Uniprot ID	Proteína	Secuencia	Sub-secuencia*	FoldX**	Categoría***
AF17_HUMAN	Protein AF-17	LLCEEEVLEVDNVKYC	LLCEEEVL	1.72	SP
CB068_HUMAN	UPF0561 protein C2orf68	LEPSGHQLFCLEYEAD	LFCLEYEA	2.45	SP
CE162_HUMAN_A	Centrosomal protein of 162 kDa	LLSTDSLETNELVVSE	LETNELVV	4.45	W
CE162_HUMAN_B	Centrosomal protein of 162 kDa	SLLSTDSLETNELVVS	LETNELVV	4.45	N
CE295_HUMAN	Centrosomal protein of 295 kDa	SLLSYENTDLSLTDPE	LLSYENTD	3.38	SP
CPSF7_HUMAN	Cleavage and polyadenylation specificity factor subunit 7	GVDLIDIYADEEFNQD	IYADEEFN	4.34	SP
DICER_HUMAN	Endoribonuclease Dicer	YSIQNLYSYENQPQPS	LYSYENQP	8.33	N
E2F4_HUMAN	Transcription factor E2F4	EELMSSEVFAPLLRLS	LMSSEVFA	03.02	W
FANCM_HUMAN	Fanconi anemia group M protein	QFLISDELLLDNNSEL	LISDELLL	1.78	W
GATA6_HUMAN[#]	Transcription factor GATA-6	GNLSSWEDLLLFTDLD	LSSWEDLL	3.12	N
GTSE1_HUMAN	G2 and S phase-expressed protein 1	DILLLADEKFDFDLSL	LLADEKFD	1.5	SP
HMBX1_HUMAN	Homeobox-containing protein 1	LHALETLDRLDQEHSD	LHALETLD	0.43	N
KDM5C_HUMAN	Lysine-specific demethylase 5C	TFLESKEELSHSPEPA	LESKEELS	2.13	N
KIF15_HUMAN	Kinesin-like protein KIF15	QELFSSERIDWTKQQE	LFSSERID	2.62	SP
KIF24_HUMAN	Kinesin-like protein KIF24	QSRETVLFSHEHMGSE	LFSHEHMG	0.22	SP
KMT2A_HUMAN	Histone-lysine N-methyltransferase 2A	LSSLESSRRVHTSTPS	LSSLESSR	6.37	SP
KPCB_HUMAN	Protein kinase C beta type	FKLLSQEEGEYFNVPV	LLSQEEGE	4.77	SP
LIN52_HUMAN	Protein lin-52 homolog	LEASLLSFEKLDRASPDL	LLSFEKLD	1.1	SP
NEF_HV1M2	Protein Nef	NADLAWLEAQEEEVGF	LEAQEEEV	4.27	W
S31D1_HUMAN	Spermatogenesis-associated protein 31D1	QQLLSWESLKDAAPSV	LLSWESLK	2.13	SP

Uniprot ID	Proteína	Secuencia	Sub-secuencia*	FoldX**	Categoría***
SPIDR_HUMAN	DNA repair-scaffolding protein	SLTSDEKLSELPKPSS	LTSDEKLS	2.34	W
TLXNB_HUMAN	Putative TLX1 neighbor protein	HSLLSQEAMGPGEGAE	LLSQEAMG	0.37	W
TTK_HUMAN	Dual specificity protein kinase TTK	LVSDEKSSELIITDSI	LVSDEKSS	5.11	SP
UIMC1_HUMAN	BRCA1-A complex subunit RAP80	VCPETQLSSSETFDLE	LSSSETFD	4.54	W
XPO4_HUMAN	Exportin-4	TNLLSKEFIDFSDTDE	LLSKEFID	2.51	W
ZN445_HUMAN	Zinc finger protein 445	WLEAREPWGLNMQAAQ	LEAREPWG	10.05	W

*Sub-secuencia en la que se detecta la expresión regular `[IL].[CAST].E`

**Valor de Foldx obtenido para la sub-secuencia al ser escaneada con 1GUX_8

***Péptidos ensayados en el laboratorio en los que se midió la afinidad de interacción; SP: Strong Positive Binder; W: Weak positive Binder; N: Negative Binder

[#] El péptido de la proteína GATA6_HUMAN fue *hit* de Rb en un cribado no analizado en el presente trabajo.

Tabla S18. Hits ProP-PD de la proteína p107 con SLiM LxCxE testeados experimentalmente escaneados con 1GUX_8.

Uniprot ID	Proteína	Secuencia	Sub-secuencia*	FoldX**	Categoría***
CB068_HUMAN	UPF0561 protein C2orf68	LEPSGHQLFCLEYEAD	LFCLEYEA	2.45	SP
CE162_HUMAN_A	Centrosomal protein of 162 kDa	LLSTDSLETNELVVSE	LETNELVV	4.45	SP
CE162_HUMAN_B	Centrosomal protein of 162 kDa	SLLSTDSLETNELVVS	LETNELVV	4.45	W
CE295_HUMAN	Centrosomal protein of 295 kDa	SLLSYENTDLSLTDPE	LLSYENTD	3.38	SP
DC8L2_HUMAN	DDB1- and CUL4-associated factor 8-like protein 2	HFLMSGESLFHYPLVG	LMSGESLF	1.32	W
DICER_HUMAN	Endoribonuclease Dicer	YSIQNLYSYENQPQPS	LYSYENQP	8.33	N
E2F4_HUMAN	Transcription factor E2F4	EELMSSEVFAPLLRLS	LMSSEVFA	03.02	W
FANCM_HUMAN	Fanconi anemia group M protein	QFLISDELLLDNNSEL	LISDELLL	1.78	W
GATA6_HUMAN	Transcription factor GATA-6	GNLSSWEDLLLFTDLD	LSSWEDLL	3.12	SP
GTSE1_HUMAN	G2 and S phase-expressed protein 1	DILLLADEKFDFDLSL	LLADEKFD	1.5	SP
HMBX1_HUMAN	Homeobox-containing protein 1	LHALETLDRLDQEHSD	LHALETLD	0.43	W
KDM5C_HUMAN	Lysine-specific demethylase 5C	TFLESKEELSHSPEPA	LESKEELS	2.13	N
KIF15_HUMAN	Kinesin-like protein KIF15	QELFSSERIDWTKQQE	LFSSERID	2.62	W
KIF24_HUMAN	Kinesin-like protein KIF24	QSRETVLFSHEHMGSE	LFSHEHMG	0.22	SP
KMT2A_HUMAN	Histone-lysine N-methyltransferase 2A	LSSLESSRRVHTSTPS	LSSLESSR	6.37	N
KPCB_HUMAN	Protein kinase C beta type	FKLLSQEEGEYFNVPV	LLSQEEGE	4.77	N
NEF_HV1M2	Protein Nef	NADLAWLEAQEEEVGF	LEAQEEEV	4.27	W
S31D1_HUMAN	Spermatogenesis-associated protein 31D1	QQLLSWESLKDAAPSV	LLSWESLK	2.13	N
SPIDR_HUMAN	DNA repair-scaffolding protein	SLTSDEKLSELPKPSS	LTSDEKLS	2.34	N
TLXNB_HUMAN	Putative TLX1 neighbor protein	HSLLSQEAMGPGEGAE	LLSQEAMG	0.37	W

Uniprot ID	Proteína	Secuencia	Sub-secuencia*	FoldX**	Categoría***
XPO4_HUMAN	Exportin-4	TNLLSKEFIDFSDTDE	LLSKEFID	2.51	W
ZN445_HUMAN	Zinc finger protein 445	WLEAREPWGLNMQAAQ	LEAREPWG	10.05	W

*Sub-secuencia en la que se detecta la expresión regular `[IL].[CAST].E`
**Valor de Foldx obtenido para la sub-secuencia al ser escaneada con 1GUX_8
***Péptidos ensayados en el laboratorio en los que se midió la afinidad de interacción; SP: Strong Positive Binder; W: Weak positive Binder; N: Negative Binder

Tabla S19. Hits ProP-PD de la proteína Rb con SLiM E2F testeados experimentalmente escaneados con 1N4M_5.

Uniprot ID	Proteína	Secuencia	Sub-secuencia*	FoldX**	Categoría***
CNOT3_HUMAN	CCR4-NOT transcription complex subunit 3	GIEDPVPTLHLTERDI	VPTLH	6	N
CTF18_HUMAN	Chromosome transmission fidelity protein 18 homolog	PEDLAELWGHGVSEAA	LAELW	5.65	W
FIP1_HUMAN	Pre-mRNA 3'-end-processing factor FIP1	PGADLSDYFNYGFNED	LSDYF	02.09	W
HFM1_HUMAN	Probable ATP-dependent DNA helicase HFM1	MLKSNDCLFSLENLFF	LENLF	2.49	W
INCE_HUMAN	Inner centromere protein	YHPPNLLELFGTILPL	LLELF	1.56	W
MYT1L_HUMAN	Myelin transcription factor 1-like protein	YVTTLTEMYTNQDRYQ	LTEMY	1.18	W
PRUN2_HUMAN	Protein prune homolog 2	SGIMELYGSDIEPQPS	IMELY	-1	W
PTF1A_HUMAN	Pancreas transcription factor 1 subunit alpha	VLLEHFPGGLDAFPSS	LLEHF	1.62	W
RAD17_HUMAN	Cell cycle checkpoint protein RAD17	GNLSSLEQIYGLENSK	LEQIY	5.48	W
SEPT7_HUMAN	Septin-7	SLFLTDLYSPEYPGPS	LTDLY	1.7	SP
SRC_HUMAN	Proto-oncogene tyrosine-protein kinase Src	ERPTFEYLQAFLEDYF	LEDYF	2.43	SP
STAR9_HUMAN	StAR-related lipid transfer protein 9	GVSDFFSTSEKEASYD	VSDFF	-0.54	SP
TFE3_HUMAN	Transcription factor E3	LLDLHFPSDHLGDLGD	LLDLH	3.12	W
XPOT_HUMAN	Exportin-T	LVELWGGKDGPVGFAD	LVELW	5.73	N

*Sub-secuencia en la que se detecta la expresión regular `[IVLMA].[NQDE][IVLFMYAW][IVLFMYAW]`

**Valor de Foldx obtenido para la sub-secuencia al ser escaneada con 1N4M_5

***Péptidos ensayados en el laboratorio en los que se midió la afinidad de interacción; SP: Strong Positive Binder; W: Weak positive Binder; N: Negative Binder

Tabla S20. Hits ProP-PD de la proteína p107 con SLiM E2F testeados experimentalmente escaneados con 1N4M_5.

Uniprot ID	Proteína	Secuencia	Sub-secuencia*	FoldX**	Categoría***
CNOT3_HUMAN	CCR4-NOT transcription complex subunit 3	GIEDPVPTLHLTERDI	VPTLH	6	N
CTF18_HUMAN	Chromosome transmission fidelity protein 18 homolog	PEDLAELWGHGVSEAA	LAELW	5.65	W
INCE_HUMAN	Inner centromere protein	YHPPNLLELFGTILPL	LLELF	1.56	W
MYT1L_HUMAN	Myelin transcription factor 1-like protein	YVTTLTEMYTNQDRYQ	LTEMY	1.18	SP
PRUN2_HUMAN	Protein prune homolog 2	SGIMELYGSDIEPQPS	IMELY	-1	W
PTF1A_HUMAN	Pancreas transcription factor 1 subunit alpha	VLLEHFPGGLDAFPSS	LLEHF	1.62	N
RAD17_HUMAN	Cell cycle checkpoint protein RAD17	GNLSSLEQIYGLENSK	LEQIY	5.48	N
SEPT7_HUMAN	Septin-7	SLFLTDLYSPEYPGPS	LTDLY	1.7	N
SRC_HUMAN	Proto-oncogene tyrosine-protein kinase Src	ERPTFEYLQAFLEDYF	LEDYF	2.43	SP
STAR9_HUMAN	StAR-related lipid transfer protein 9	GVSDFFSTSEKEASYD	VSDFF	-0.54	N
TFE3_HUMAN	Transcription factor E3	LLDLHFPSDHLGDLGD	LLDLH	3.12	N
XPOT_HUMAN	Exportin-T	LVELWGGKDGPVGFAD	LVELW	5.73	N

*Sub-secuencia en la que se detecta la expresión regular `[IVLMA].[NQDE][IVLFMYAW][IVLFMYAW]`
**Valor de Foldx obtenido para la sub-secuencia al ser escaneada con 1N4M_5
***Péptidos ensayados en el laboratorio en los que se midió la afinidad de interacción; SP: Strong Positive Binder; W: Weak positive Binder; N: Negative Binder

Tabla S21. Criterios de priorización aplicados a péptidos *hit* de Rb con patrones de secuencia del SLiM LxCxE.

Uniprot ID	Proteína	Secuencia[1]	Variante[#]	RSA[##]	IUPred[##]	Pfam* [##]	Localización Celular**	FoldX***	Testeado [†]	Criterio aplicado
CPSF7_HUMAN	Cleavage and polyadenylation specificity factor subunit 7	GVDLI**DIYADEEF**NQD	1	0,83	0,46	0	N	4,34	SP	Priorizable
KIF24_HUMAN	Kinesin-like protein KIF24	QSRETV**LFSHEH**MGSE	2A	0,81	0,62	0	C	0,22	SP	Priorizable (Localización)
HMBX1_HUMAN	Homeobox-containing protein 1	**LHALETL**DRLDQEHSD	2A	0,52	0,65	16	NC	0,43	N	Priorizable (Pfam)
RBM33_HUMAN	RNA-binding protein 33	EEQ**LYTDEVL**DIEINE	2A	0,84	0,72	16	NC	0,84	**No testeado**	Priorizable (Pfam)
HFM1_HUMAN	Probable ATP-dependent DNA helicase HFM1	MLKSNDC**LFSLENL**FF	2A	0,88	0,2	0	NC	1,43	W	Priorizable (IUPred)
HFM1_HUMAN	Probable ATP-dependent DNA helicase HFM1	**LFSLENL**FFEKPDEVE	2A	0,88	0,38	0	NC	1,43	W	Priorizable (IUPred)
KDM5A_HUMAN[%]	Lysine-specific demethylase 5A	EPN**LFCDEEI**PIKSEE	2A	0,85	0,42	0	N	4,09	SP (LxCxE TP)	Priorizable
SUSD6_HUMAN	Sushi domain-containing protein 6	A**LPSYEEA**VYGSSGHC	2A	0,64	0,36	0	C	5,81	**No testeado**	Priorizable (IUPred, Localización, FoldX)
TAF1_HUMAN	Transcription initiation factor TFIID subunit 1	SLIT**ELTANEEL**TGTD	2B	0,77	0,5	11	N	0,93	**No testeado**	Priorizable (Pfam)
SIK3_HUMAN	Serine/threonine-protein kinase SIK3	PLNEDV**LLAMEDM**GLD	3	0,4	0,38	0	C	-1,45	No testeado	Priorizable (IUPred, Localización)
VWA3B_HUMAN	von Willebrand factor A domain-containing protein 3B	MSI**LLAEEWL**DDKSSE	3	0,45	0,38	0	C	-0,81	No testeado	Priorizable (IUPred, Localización)
VWA3B_HUMAN	von Willebrand factor A domain-containing protein 3B	SDKEMSI**LLAEEWL**DD	3	0,45	0,37	0	C	-0,81	No testeado	Priorizable (IUPred, Localización)
CROCC_HUMAN	Rootletin	QAL**LLAKETL**TGELAG	3	0,53	0,37	0	C	-0,34	No testeado	Priorizable (IUPred, Localización)
SPKAP_HUMAN	A-kinase anchor protein SPHKAP	DF**LTASEHL**EEESEVD	3	0,75	0,51	0	C	1,04	No testeado	Priorizable (Localización)
CAC1C_HUMAN	Voltage-dependent L-type calcium channel subunit alpha-1C	SEPS**LLSTEM**LSYQDD	3	0,82	0,56	0	C	1,11	No testeado	Priorizable (Localización)
GTSE1_HUMAN	G2 and S phase-expressed protein 1	DIL**LLADEKF**DFDLSL	3	0,75	0,33	15	NC	1,5	SP	Priorizable (IUPred, Pfam)
FANCM_HUMAN	Fanconi anemia group M protein	QF**LISDEL**LLDNNSEL	3	0,83	0,37	0	N	1,78	W	Priorizable (IUPred)
PCNT_HUMAN	Pericentrin	SVQK**LLAAEQTV**VRDL	3	0,52	0,38	0	C	2,09	**No testeado**	Priorizable (IUPred, Localización)

Uniprot ID	Proteína	Secuencia[1]	Variante[#]	RSA[##]	IUPred[##]	Pfam* [##]	Localización Celular**	FoldX***	Testeado [†]	Criterio aplicado
S31D1_HUMAN	Spermatogenesis-associated protein 31D1	QQ**LLSWESL**KDAAPSV	3	0,73	0,41	16	C	2,13	SP	Priorizable (Pfam, Localización)
E2F2_HUMAN	Transcription factor E2F2	G**LEAGEGI**SDLFDSYD	3	0,67	0,38	0	N	2,13	SP (E2F TP)	Priorizable (IUPred)
E2F2_HUMAN	Transcription factor E2F2	DYLWG**LEAGEGI**SDLF	3	0,65	0,42	0	N	2,13	SP (E2F TP)	Priorizable
ARHGC_HUMAN	Rho guanine nucleotide exchange factor 12	ETDPPNWQQ**LVSREVL**	3	0,45	0,46	0	C	2,48	No testeado	Priorizable (Localización)
NPHP4_HUMAN	Nephrocystin-4	FQFS**LGSEEHL**DAPTE	3	0,65	0,60	3	NC	2,50	**No testeado**	Priorizable
ZN180_HUMAN	Zinc finger protein 180	T**LLCLEESM**EEQDEKP	3	0,90	0,69	0	N	2,85	No testeado	Priorizable
SEPT7_HUMAN	Septin-7	SLFLTD**LYSPEY**PGPS	3	0,46	0,26	16	NC	4,27	SP	Priorizable (IUPred, Pfam)
RAD17_HUMAN	Cell cycle checkpoint protein RAD17	RGN**LSSLEQI**YGLENS	3	0,85	0,39	0	N	4,47	No testeado	Priorizable (IUPred)
CEBPE_HUMAN	CCAAT/enhancer-binding protein epsilon	IDLSAY**IESGEEQL**LS	3	0,62	0,45	0	N	7,81	No testeado	Priorizable (FoldX)
ZN445_HUMAN	Zinc finger protein 445	W**LEAREPW**GLNMQAAQ	3	0,62	0,51	0	N	10,05	W	Priorizable (FoldX)
CE295_HUMAN	Centrosomal protein of 295 kDa	S**LLSYE**NTDLSLTDPE	4B	0,76	0,51	0	C	3,38	SP	Priorizable (Localización)
GASP1_HUMAN	G-protein coupled receptor-associated sorting protein 1	IGSW**LWATE**ESNIDGT	5	0,81	0,44	0	C	3,23	No testeado	Priorizable (Localización)
GASP1_HUMAN	G-protein coupled receptor-associated sorting protein 1	PEAIIGSW**LWATE**ESN	5	0,81	0,37	0	C	3,23	No testeado	Priorizable (IUPred, Localización)
ZN184_HUMAN	Zinc finger protein 184	WSSNL**LESWE**YEGSLE	5	0,85	0,45	0	N	5,07	No testeado	Priorizable (FoldX)
ZN184_HUMAN	Zinc finger protein 184	L**LESWE**YEGSLERQQA	5	0,83	0,49	0	N	5,07	No testeado	Priorizable (FoldX)
SGSM1_HUMAN	Small G protein signaling modulator 1	SLES**DLLANESM**DEFM	2B	0,93	0,61	0	SE	-0,97	No testeado	No priorizable (Localización)
SGSM1_HUMAN	Small G protein signaling modulator 1	**DLLANESM**DEFMSITG	2B	0,92	0,58	0	SE	-0,97	No testeado	No priorizable (Localización)
TEX2_HUMAN	Testis-expressed protein 2	G**LEAKEDL**YLEPQVGH	3	0,75	0,63	16	SE	0,94	No testeado	No priorizable (Localización)
ZDH12_HUMAN	Palmitoyltransferase ZDHHC12	WET**LWAEE**EEEGSSPA	5	0,45	0,3	0	SE	2,57	No testeado	No priorizable (Localización)

Uniprot ID	Proteína	Secuencia[1]	Variante[#]	RSA[##]	IUPred[##]	Pfam* [##]	Localización Celular**	FoldX***	Testeado [†]	Criterio aplicado
ASB3_HUMAN	Ankyrin repeat and SOCS box protein 3	GADP**DLYCNEDSW**QLP	1	0,26	0,11	8	NC	3,31	No testeado	Filtrado (RSA, IUPred, Pfam)
ABCA8_HUMAN	ATP-binding cassette sub-family A member 8	SHLLFSS**LFSEERM**DV	2A	0,56	0,01	16	C	0,28	No testeado	Filtrado (IUPred, Pfam)
NCKX1_HUMAN	Sodium/potassium/calcium exchanger 1	S**LFSREIL**NLTWWPLF	2A	0,31	0,01	16	C	1,59	No testeado	Filtrado (RSA, IUPred, Pfam)
RP1L1_HUMAN	Retinitis pigmentosa 1-like 1 protein	HS**LSALEQL**EDGGCYL	3	0,38	0,42	16	C	1,14	No testeado	Filtrado (RSA)
AF17_HUMAN	Protein AF-17	**LLCEEEV**LEVDNVKYC	3	0,48	0,1	16	N	1,72	SP	Filtrado (IUPred, Pfam)
XPO1_HUMAN	Exportin-1	**LLSEEVF**DFSSGQITQ	3	0,22	0,07	16	NC, SE	2,78	No testeado	Filtrado (RSA, IUPred, Pfam)
EPHA1_HUMAN	Ephrin type-A receptor 1	PYVD**LQAYEDPA**QGAL	3	0,39	0,24	16	C	3,89	No testeado	Filtrado (RSA)

[1]Se destaca en negrita la variante de expresión regular y en subrayado la subsecuencia a la cual corresponde el valor de FoldX de cada péptido

[#]Variante 1: `[DE][IL][YFH][CAST].E.{1,2}[WFILYVM]o [DE][IL].[CAST][YFH]E.{1,2}[WFILYVM];`

Variante 2A: `[IL][YFH][CAST].E.{1,2}[WFILYVM]o [IL].[CAST][YFH]E.{1,2}[WFILYVM];`

Variante 2B: `[DE][IL].[CAST].E.{1,2}[WFILYVM];`

Variante 3: `[IL].[CAST].E.{1,2}[WFILYVM]`

Variante 4A: `[DE][IL].[CAST].E`

Variante 5: `[IL].[CAST].E`

[##]Celdas en amarillo indican una advertencia. Celdas en rojo señalan el motivo de filtrado.

* Número de residuos solapados con dominios Pfam

** N= Núcleo; NC= Núcleo y Citoplasma, C= Citoplasma; SE= Sistema Endomembrana

***Valor mínimo de secuencia escaneada con matriz FoldX 1GUX_8.

[†]SP= Péptido de interacción *Strong Positive* de acuerdo a ensayos del laboratorio, W = Péptido de interacción *Weak* de acuerdo a ensayos del laboratorio, N= Péptido sin interacción de acuerdo a ensayos del laboratorio, E2F TP = Péptido conteniendo al SLiM E2F reportado en ELM como verdadero positivo, LxCxE TP = Péptido conteniendo al SLiM LxCxE reportado en ELM como verdadero positivo

[%]El péptido de KDM5A_HUMAN presenta dos Variantes de regex (2A, siete posiciones en negrita correspondiente al SLiM funcional y 5, cinco posiciones subrayadas). El mínimo valor de FoldX corresponde a la Variante 5, donde la matriz incluye sólo las 5 posiciones subrayadas. La variante 2A corresponde al siguiente mínimo de FoldX (Subsecuencia: `LFCDEEIP` ; FoldX= 4,52)

.

Tabla S22. Criterios de priorización aplicados a péptidos *hit* de Rb con patrones de secuencia del SLiM E2F.

Uniprot ID	Proteína	Secuencia[1]	Variante [#]	RSA[##]	IUPred[##]	Pfam* [##]	Localización Celular**	FoldX***	Testeado [†]	Criterio aplicado
HAX1_HUMAN	HCLS1-associated protein X-1	PALDDAFS**ILDLF**LGR	1	0,65	0,35	15	NC, SE	-1,64	**No testeado**	Priorizable (IUpred y Pfam)
PRUN2_HUMAN	Protein prune homolog 2	SG**IMELY**GSDIEPQPS	1	0,82	0,53	0	C	-1	W	Priorizable (Localización)
AGGF1_HUMAN	Angiogenic factor with G patch and FHA domains 1	EVQTENHAPWS**ISDYF**	1	0,81	0,51	0	C	-0,63	**No testeado**	Priorizable (Localización)
STAR9_HUMAN	StAR-related lipid transfer protein 9	G**VSDFF**STSEKEASYD	1	0,87	0,41	0	NC	-0,54	SP	Priorizable
STAR9_HUMAN	StAR-related lipid transfer protein 9	DSLNAKLEG**VSDFF**ST	1	0,84	0,44	0	NC	-0,54	SP	Priorizable
GON7_HUMAN	EKC/KEOPS complex subunit GON7	**VTELF**DPLVQGEVQHR	1	0,55	0,63	16	N	-0,31	**No testeado**	Priorizable (Pfam)
E2F1_HUMAN	Transcription factor E2F1	DYHFGLEEGEG**IRDLF**	1	0,69	0,4	0	NC	-0,1	SP (E2F TP)	Priorizable
E2F5_HUMAN	Transcription factor E2F5	YNFNLDDNEG**VCDLF**D	1	0,65	0,4	0	NC	-0,04	SP (E2F TP)	Priorizable
E2F2_HUMAN	Transcription factor E2F2	GLEAGEG**ISDLF**DSYD	1	0,67	0,38	0	N	0	SP (E2F TP)	Priorizable (IUPred)
E2F2_HUMAN	Transcription factor E2F2	GEG**ISDLF**DSYDLGDL	1	0,66	0,32	0	N	0	SP (E2F TP)	Priorizable (IUPred)
E2F2_HUMAN	Transcription factor E2F2	DYLWGLEAGEG**ISDLF**	1	0,65	0,42	0	N	0	SP (E2F TP)	Priorizable
E2F3_HUMAN	Transcription factor E2F3	DYLLSLGEEEG**ISDLF**	1	0,65	0,39	0	NC	0	SP (E2F TP)	Priorizable (IUPred)
PKHG2_HUMAN	Pleckstrin homology domain-containing family G member 2	**ISDVF**EMPCLPAIPSV	1	0,81	0,54	0	C	0,98	No testeado	Priorizable (Localización)
SCAF8_HUMAN	SR-related and CTD-associated factor 8	**VFDYF**EGATSQRKGDN	1	0,86	0,68	0	N	1,16	**No testeado**	Priorizable
MYT1L_HUMAN	Myelin transcription factor 1-like protein	YVTT**LTEMY**TNQDRYQ	1	0,41	0,54	0	N	1,18	W	Priorizable
SYDE2_HUMAN	Rho GTPase-activating protein SYDE2	VLSVPPDQRIT**LTDLF**	1	0,79	0,44	0	C	1,54	No testeado	Priorizable (Localización)

Uniprot ID	Proteína	Secuencia[1]	Variante [#]	RSA[##]	IUPred[##]	Pfam* [##]	Localización Celular**	FoldX***	Testeado [†]	Criterio aplicado
INCE_HUMAN	Inner centromere protein	YHPPN**LLELF**GTILPL	1	0,63	0,27	16	NC	1,56	W	Priorizable (IUPred, Pfam)
SEPT7_HUMAN	Septin-7	SLF**LTDLY**SPEYPGPS	1	0,46	0,26	16	NC	1,7	SP	Priorizable (IUPred, Pfam)
VPS54_HUMAN	Vacuolar protein sorting-associated protein 54	GMF**ISDAF**GEGELTPI	1	0,86	0,28	0	NC, SE	1,72	**No testeado**	Priorizable (IUpred)
VPS54_HUMAN	Vacuolar protein sorting-associated protein 54	LIHEGMF**ISDAF**GEGE	1	0,84	0,21	0	NC, SE	1,72	No testeado	Priorizable (IUpred)
VPS54_HUMAN	Vacuolar protein sorting-associated protein 54	EVAYLIHEGMF**ISDAF**	1	0,81	0,19	0	NC, SE	1,72	No testeado	Priorizable (IUpred)
EPN3_HUMAN	Epsin-3	SQSS**ILDLA**DIFVPAL	1	0,75	0,49	0	NC, SE	1,87	No testeado	Priorizable
HUWE1_HUMAN	E3 ubiquitin-protein ligase HUWE1	L**LDDFF**HDQSTATSQA	1	0,61	0,48	0	NC, SE	2,49	No testeado	Priorizable
CDR2_HUMAN	Cerebellar degeneration-related protein 2	KEPSQSL**LEEMF**LTVP	1	0,68	0,4	0	C	2,54	No testeado	Priorizable (Localización)
RHG17_HUMAN	Rho GTPase-activating protein 17	DWFFPEEVEFN**VSEAF**	1	0,43	0,22	0	NC	2,58	No testeado	Priorizable (IUpred)
EPN4_HUMAN	Clathrin interactor 1	**LVDLF**DGTSQSTGGSA	1	0,91	0,56	0	NC, SE	2,69	No testeado	Priorizable
I20RA_HUMAN	Interleukin-20 receptor subunit alpha	LGYASH**LMEIF**CDSEE	1	0,86	0,38	0	C	2,85	No testeado	Priorizable (IUPred, Localización)
EM55_HUMAN	55 kDa erythrocyte membrane protein	GSMHTA**LSDLY**LEHLL	1	0,57	0,46	0	C	2,88	No testeado	Priorizable (Localización)
CPSF7_HUMAN	Cleavage and polyadenylation specificity factor subunit 7	GVD**LIDIY**ADEEFNQD	1	0,83	0,46	0	N	3,71	SP	Priorizable (FoldX)
NRK_HUMAN	Nik-related protein kinase	NA**LSEIF**RNDWLTPAP	1	0,79	0,42	0	C	4,49	No testeado	Priorizable (Localización, FoldX)
ZN317_HUMAN	Zinc finger protein 317	H**LFEVF**GMDPHLTQPM	1	0,73	0,46	0	N	5,58	No testeado	Priorizable (FoldX)
HFM1_HUMAN	Probable ATP-dependent DNA helicase HFM1	MLKSNDCLFS**LENLF**F	2	0,88	0,2	0	NC	2,49	W	Priorizable (IUPred)

Uniprot ID	Proteína	Secuencia[1]	Variante [#]	RSA[##]	IUPred[##]	Pfam* [##]	Localización Celular**	FoldX***	Testeado [†]	Criterio aplicado
RAD17_HUMAN	Cell cycle checkpoint protein RAD17	RGNLSS**LEQIY**GLENS	2	0,85	0,39	0	N	5,48	W	Priorizable (IUPred, FoldX)
RAD17_HUMAN	Cell cycle checkpoint protein RAD17	SS**LEQIY**GLENSKEYL	2	0,81	0,37	3	N	5,48	W	Priorizable (IUPred, FoldX)
ETAA1_HUMAN	Ewing's tumor-associated antigen 1	**MLDMW**IGETAIPCTPS	3	0,71	0,48	16	N	-0,11	No testeado	Priorizable (Pfam)
BMAL1_HUMAN	Aryl hydrocarbon receptor nuclear translocator-like protein 1	**ISDFM**SPGPTDLLSSS	3	0,89	0,58	0	NC	-0,04	No testeado	Priorizable
VWA3B_HUMAN	von Willebrand factor A domain-containing protein 3B	SESL**IMDWW**YNAEKDG	3	0,45	0,29	0	C	0,52	No testeado	Priorizable (IUPred, Localización)
VWA3B_HUMAN	von Willebrand factor A domain-containing protein 3B	**IMDWW**YNAEKDGDSKH	3	0,45	0,33	0	C	0,52	No testeado	Priorizable (IUPred, Localización)
MLF2_HUMAN	Myeloid leukemia factor 2	MSGGFMDMFGM**MNDMI**	3	0,63	0,28	16	NC	1,77	No testeado	Priorizable (IUpred y Pfam)
PTPA_HUMAN	Serine/threonine-protein phosphatase 2A activator	**VSEMW**NEVHEEKEQAA	3	0,77	0,47	16	NC	2,44	No testeado	Priorizable (Pfam)
BCL9_HUMAN	B-cell CLL/lymphoma 9 protein	S**LQDMM**VHQHGPRGVV	3	0,69	0,76	0	NC, SE	2,72	No testeado	Priorizable
TPM3_HUMAN	Tropomyosin alpha-3 chain	**MMEAI**KKKMQMLKLDK	3	0,68	0,48	0	C	3,05	No testeado	Priorizable (Localización, FoldX)
SPNDC_HUMAN	Uncharacterized protein C11orf84	**ILEAW**SEGVALLQDVR	3	0,43	0,63	7	NC	3,12	No testeado	Priorizable (FoldX)
CP4Z2_HUMAN	Putative inactive cytochrome P450 family member 4Z2	MEPSW**LQELM**AHPFLL	3	0,59	0,01	0	C, SE	4,2	No testeado	Priorizable (IUPred, FoldX, Localización)
VWA3B_HUMAN	von Willebrand factor A domain-containing protein 3B	SDKEMSILL**AEEWL**DD	3	0,45	0,37	0	C	4,51	No testeado	Priorizable (IUPred, FoldX, Localización)
VWA3B_HUMAN	von Willebrand factor A domain-containing protein 3B	MSILL**AEEWL**DDKSSE	3	0,45	0,38	0	C	4,51	No testeado	Priorizable (IUPred, FoldX, Localización)

Uniprot ID	Proteína	Secuencia[1]	Variante [#]	RSA[##]	IUPred[##]	Pfam* [##]	Localización Celular**	FoldX***	Testeado [†]	Criterio aplicado
DREB_HUMAN	Drebrin	S**LIDLW**PGNGEGASTL	3	0,88	0,82	0	C	4,83	No testeado	Priorizable (Localización, FoldX)
KMT2D_HUMAN	Histone-lysine N-methyltransferase 2D	H**LDDLL**NGDEFDLLAY	3	0,69	0,66	0	N	4,88	No testeado	Priorizable (FoldX)
PKHG2_HUMAN	Pleckstrin homology domain-containing family G member 2	LPLPQV**LTDIW**VQALP	3	0,77	0,54	0	C	5,4	No testeado	Priorizable (FoldX, Localización)
PKHG2_HUMAN	Pleckstrin homology domain-containing family G member 2	ATTPLPLPQV**LTDIW**V	3	0,76	0,59	0	C	5,4	No testeado	Priorizable (FoldX, Localización)
CTF18_HUMAN	Chromosome transmission fidelity protein 18 homolog	PED**LAELW**GHGVSEAA	3	0,87	0,61	0	NC	5,65	W	Priorizable (FoldX)
DDIT3_HUMAN	DNA damage-inducible transcript 3 protein	WELEAWYED**LQEVL**SS	3	0,55	0,5	0	NC, SE	5,77	No testeado	Priorizable (FoldX)
CLCKB_HUMAN	Chloride channel protein ClC-Kb	REGSSGNPVT**LQELW**G	3	0,85	0,23	0	C	5,88	No testeado	Priorizable (IUPred, FoldX, Localización)
TCHP_HUMAN	Trichoplein keratin filament-binding protein	EEEEEARRVEQ**LSDAL**	3	0,57	0,63	16	C	5,9	No testeado	Priorizable (Pfam, Localización, FoldX)
S31D1_HUMAN	Spermatogenesis-associated protein 31D1	QQ**LLSWESLK**DAAPSV	3	0,73	0,41	16	C	8,41	SP	Priorizable (Pfam, Localización, FoldX)
MPIP3_HUMAN	M-phase inducer phosphatase 3	**MLNLL**LERDTSFTVCP	4	0,68	0,4	0	NC	-0,78	No testeado	Priorizable
ZN506_HUMAN	Zinc finger protein 506	IVKN**VENLL**NVPQPLI	4	0,82	0,24	0	N	1,62	No testeado	Priorizable (IUpred)
I20RA_HUMAN	Interleukin-20 receptor subunit alpha	GENETY**LMQFM**EEWGL	4	0,62	0,59	0	C	2,01	No testeado	Priorizable (Localización)
MMAD_HUMAN	Methylmalonic aciduria and homocystinuria type D protein, mitochondrial	V**MAQYV**NEFQGNDAPV	4	0,64	0,4	16	C	2,74	No testeado	Priorizable (Pfam, Localización)
GTPBA_HUMAN	GTP-binding protein 10	Q**LLNLW**ISDTMSSTEP	4	0,59	0,43	0	NC	3,08	No testeado	Priorizable (FoldX)

Uniprot ID	Proteína	Secuencia[1]	Variante [#]	RSA[##]	IUPred[##]	Pfam* [##]	Localización Celular**	FoldX***	Testeado [†]	Criterio aplicado
F131B_HUMAN	Protein FAM131B	GV**MEQFA**ISEATLMAW	4	0,5	0,29	16	NC	3,86	No testeado	Priorizable (IUpred, Pfam, FoldX)
FOXP2_HUMAN	Forkhead box protein P2	**LLQAV**HEDLNGSLDHI	4	0,8	0,52	0	N	6,32	No testeado	Priorizable (FoldX)
RN216_HUMAN	E3 ubiquitin-protein ligase RNF216	QEPN**LENIW**GQEAAEV	4	0,85	0,57	0	NC	6,35	No testeado	Priorizable (FoldX)
MYRF_HUMAN	Myelin regulatory factor	M**LHQLL**QQHGAELPTH	4	0,86	0,9	0	NC, SE	6,52	No testeado	Priorizable (FoldX)
EF1D_HUMAN	Elongation factor 1-delta	E**LQQAI**SKLEARLNVL	4	0,56	0,45	0	NC	8,06	No testeado	Priorizable (FoldX)
CLC14_HUMAN	C-type lectin domain family 14 member A	MRPAFALCL**LWQAL**WP	4	0,74	0,07	0	C	8,37	No testeado	Priorizable (IUpred, Localización, FoldX)
AP180_HUMAN	Clathrin coat assembly protein AP180	AKVDSSG**VIDLF**GDAF	1	0,87	0,44	0	SE	-0,07	No testeado	No priorizable (Localización)
AP180_HUMAN	Clathrin coat assembly protein AP180	SSG**VIDLF**GDAFGSSA	1	0,87	0,45	0	SE	-0,07	No testeado	No priorizable (Localización)
GRIP1_HUMAN	Glutamate receptor-interacting protein 1	**LSDMY**PSTVPSVDSAV	1	0,86	0,56	0	SE	1,88	No testeado	No priorizable (Localización)
CC190_HUMAN	Coiled-coil domain-containing protein 190	LS**IGEIF**GHGESSSSR	1	0,71	0,57	7	O	2,26	No testeado	No Priorizable (Localización)
CC190_HUMAN	Coiled-coil domain-containing protein 190	SERLLS**IGEIF**GHGES	1	0,65	0,55	11	O	2,26	No testeado	No Priorizable (Localización)
TBC9B_HUMAN	TBC1 domain family member 9B	**LYNMF**SEDPMEQDLYH	1	0,41	0,45	0	SE	2,45	No testeado	No Priorizable (Localización)
SGSM1_HUMAN	Small G protein signaling modulator 1	SLESDLLANES**MDEFM**	3	0,93	0,61	0	SE	1,09	No testeado	No priorizable (Localización)
SGSM1_HUMAN	Small G protein signaling modulator 1	DLLANES**MDEFM**SITG	3	0,92	0,58	0	SE	1,09	No testeado	No priorizable (Localización)
BRAS2_HUMAN	Putative uncharacterized protein encoded by BRWD1-AS2	I**LCNFL**PGCWLVGDVA	3	0,78	0,27	16	O	2,26	No testeado	No priorizable (Localización)
CRBG1_HUMAN	Beta/gamma crystallin domain-containing protein 1	ELSGLWG**IEDIL**ERHE	3	0,63	0,26	13	O	3,09	No testeado	No priorizable (Localización)

Uniprot ID	Proteína	Secuencia[1]	Variante [#]	RSA[##]	IUPred[##]	Pfam* [##]	Localización Celular**	FoldX***	Testeado [†]	Criterio aplicado
CRBG1_HUMAN	Beta/gamma crystallin domain-containing protein 1	GELELSGLWG**IEDIL**E	3	0,54	0,25	16	O	3,09	No testeado	No priorizable (Localización)
YK042_HUMAN	Putative uncharacterized protein ENSP00000347057	G**ILEAW**GSCGRWCGVG	3	0,9	0,25	16	O	3,12	No testeado	No priorizable (Localización)
YK042_HUMAN	Putative uncharacterized protein ENSP00000347057	VESVPEEG**ILEAW**GSC	3	0,89	0,46	0	O	3,12	No testeado	No priorizable (Localización)
MBOA1_HUMAN	Lysophospholipid acyltransferase 1	TYLHP**LSELL**GIPLDQ	3	0,41	0,01	1	SE	4,66	No testeado	No priorizable (Localización)
AKAP6_HUMAN	A-kinase anchor protein 6	**VVEAW**YGSDEYLALPS	3	0,75	0,39	0	SE	5,12	W	No priorizable (Localización)
PR15A_HUMAN	Protein phosphatase 1 regulatory subunit 15A	S**LPEAW**GLLDDDDGMY	3	0,87	0,61	0	SE	8,72	No testeado	No priorizable (Localización)
ZFTA_HUMAN	Uncharacterized protein C11orf95	**LLQAW**GGQPEALSELT	4	0,45	0,69	7	O	6,34	No testeado	No priorizable (Localización)
STML1_HUMAN	Stomatin-like protein 1	DST**LQQLA**LHFLGGSM	4	0,66	0,36	0	SE	7,32	No testeado	No priorizable (Localización)
CAC1B_HUMAN	Voltage-dependent N-type calcium channel subunit alpha-1B	G**VLNYF**RDAWNVFDFV	1	0,29	0	16	C	-2,45	No testeado	Filtrado (RSA, IUPred, Pfam)
NALCN_HUMAN	Sodium leak channel non-selective protein	TPTAVIRDFGG**VMDIF**	1	0,26	0,01	16	C	0,04	No testeado	Filtrado (RSA, IUPred, Pfam)
ERO1B_HUMAN	ERO1-like protein beta	FERS**IVDLY**TGNAEED	1	0,32	0,17	16	SE	0,12	No testeado	Filtrado (RSA, IUPred, Pfam)
ARB2P_HUMAN	Putative protein FAM172B	TAY**IWDYF**ISKTEGKD	1	0,24	0,11	16	C	0,87	No testeado	Filtrado (RSA, IUPred, Pfam)
EMC2_HUMAN	ER membrane protein complex subunit 2	**VSELY**DVTWEEMRDKM	1	0,39	0,19	0	NC, SE	1,02	No testeado	Filtrado (RSA)
SEPT2_HUMAN	Septin-2	SLF**LTDLY**PERVIPGA	1	0,41	0,15	16	NC	1,7	No testeado	Filtrado (IUPred, Pfam)

Uniprot ID	Proteína	Secuencia[1]	Variante [#]	RSA[##]	IUPred[##]	Pfam* [##]	Localización Celular**	FoldX***	Testeado [†]	Criterio aplicado
TPC12_HUMAN	Trafficking protein particle complex subunit 12	**ASDFF**DSFTTSAFISV	1	0,23	0,41	0	NC	1,72	No testeado	Filtrado (RSA)
DYH10_HUMAN	Dynein heavy chain 10, axonemal	DIIL**LSEMF**SDNFGQL	1	0,3	0,13	0	C	2,2	No testeado	Filtrado (RSA)
DYH10_HUMAN	Dynein heavy chain 10, axonemal	QGWEDIIL**LSEMF**SDN	1	0,26	0,07	0	C	2,2	No testeado	Filtrado (RSA)
KCNB2_HUMAN	Potassium voltage-gated channel subfamily B member 2	LDYWG**IDEIY**LESCCQ	1	0,38	0,06	14	C	2,62	No testeado	Filtrado (RSA, IUPred, Pfam)
GEM_HUMAN	GTP-binding protein GEM	**LLDMW**ENKGENEWLHD	3	0,37	0,26	16	NC	2,65	No testeado	Filtrado (RSA)
PA24D_HUMAN	Cytosolic phospholipase A2 delta	N**LLDAW**YDLTSSGESW	3	0,36	0,11	16	C	5,37	No testeado	Filtrado (RSA, IUPred, Pfam)
XPOT_HUMAN	Exportin-T	**LVELW**GGKDGPVGFAD	3	0,39	0,04	16	NC	5,73	N	Filtrado (RSA, IUPred, Pfam)
CERS3_HUMAN	Ceramide synthase 3	YD**LWEVW**NGYPKQPLL	3	0,29	0,01	16	SE	8,25	No testeado	Filtrado (RSA, IUPred, Pfam)
ZNG1C_HUMAN	COBW domain-containing protein 3	**LSNVL**DLHAFDSLSGI	4	0,23	0,14	0	NC, C, O	4,6	No testeado	Filtrado (RSA)
SCMC1_HUMAN	Calcium-binding mitochondrial carrier protein SCaMC-1	ELLKSYW**LDNFA**KDSV	4	0,42	0,03	11	C	5,43	No testeado	Filtrado (IUpred, Pfam)
MILK2_HUMAN	MICAL-like protein 2	L**LEQYV**STVNDRSDIV	4	0,3	0,38	16	SE	5,95	No testeado	Filtrado (RSA)
DYHC1_HUMAN	Cytoplasmic dynein 1 heavy chain 1	VNWVVSELT**LGQIW**DV	4	0,33	0,04	16	SE	8,04	No testeado	Filtrado (RSA, IUPred, Pfam)

[1]Se destaca en negrita la variante de expresión regular y en subrayado la subsecuencia a la cual corresponde el valor de FoldX de cada péptido

[#] Variante 1: `[IVLMA].[DE][IVLMAFYW][FY]`

Variante 2: `[IVLMA].[NQDE][IVLMAFYW][FY]`

Variante 3: `[IVLMA].[DE][IVLMAFYW][IVLMAFYW]`

Variante 4: `[IVLMA].[NQDE][IVLMAFYW][IVLMAFYW]`

[##]Celdas en amarillo indican una advertencia. Celdas en rojo señalan el motivo de filtrado

* Número de residuos solapados con dominios Pfam

** N= Núcleo; NC= Núcleo y Citoplasma, C= Citoplasma; SE= Sistema Endomembrana; O= Otro

***Valor mínimo de secuencia escaneada con matriz FoldX 1N4M_5.

[†]SP= Péptido de interacción *Strong Positive* de acuerdo a ensayos del laboratorio, W = Péptido de interacción *Weak* de acuerdo a ensayos del laboratorio, N= Péptido sin interacción de acuerdo a ensayos del laboratorio, E2F TP = Péptido conteniendo al SLiM E2F reportado en ELM como verdadero positivo

Tabla S23. Criterios de priorización aplicados a péptidos *hit* de p107 con patrones de secuencia del SLiM LxCxE.

Uniprot ID	Proteína	Secuencia[1]	Variante[#]	RSA[##]	IUPred[##]	Pfam* [##]	Localización Celular**	FoldX***	Testeado [†]	Criterio aplicado
KIF15_HUMAN	Kinesin-like protein KIF15	Q**ELFSSERI**DWTKQQE	1	0,6	0,4	0	C	2,62	W	Priorizable (Localización)
KIF15_HUMAN	Kinesin-like protein KIF15	STQMQ**ELFSSERI**DWT	1	0,57	0,46	0	C	2,62	W	Priorizable (Localización)
UBP10_HUMAN	Ubiquitin carboxyl-terminal hydrolase 10	GTATNGV**ELHTTESI**D	1	0,82	0,8	0	NC, SE	3,44	W	Priorizable
KIF24_HUMAN	Kinesin-like protein KIF24	QSRETV**LFSHEHM**GSE	2A	0,81	0,62	0	C	0,22	SP	Priorizable (Localización)
LIN52_HUMAN	Protein lin-52 homolog	TDLEAS**LLSFEKL**DRA	2A	0,69	0,38	10	N	1,1	SP (LxSxE TP)	Priorizable (IUPred, Pfam)
LDB3_HUMAN	LIM domain-binding protein 3	QYNNPIG**LYSAETL**RE	2A	0,63	0,57	16	C	1,82	**No testeado**	Priorizable (Pfam, Localización)
PDLI1_HUMAN	PDZ and LIM domain protein 1	G**LYSSENI**SNFNNALE	2A	0,58	0,51	16	C	3,23	**No testeado**	Priorizable (Pfam, Localización)
UTP25_HUMAN	Digestive organ expansion factor homolog	S**LFSLETNF**LEEESGD	2A	0,57	0,63	0	N	3,63	**No testeado**	Priorizable
KDM5A_HUMAN[%]	Lysine-specific demethylase 5A	EPN**LFCDEEI**PIKSEE	2A	0,85	0,42	0	N	4,09	SP (LxCxE TP)	Priorizable
CAC1F_HUMAN	Voltage-dependent L-type calcium channel subunit alpha-1F	SS**LYSDEESI**LSRFDE	2A	0,9	0,37	16	C	4,79	No testeado	Priorizable (IUPred, Pfam, Localización)
HMBX1_HUMAN	Homeobox-containing protein 1	**LHALETL**DRLDQEHSD	2A	0,52	0,65	16	NC	0,43	W	Priorizable (Pfam)
MNAR1_HUMAN	UPF0258 protein KIAA1024	KLTAL**DLQTQESL**NPN	2B	0,67	0,45	16	C	2,35	**No testeado**	Priorizable (Pfam, Localización)
ZN436_HUMAN	Zinc finger protein 436	QWG**DLTAEEW**VSYPLQ	2B	0,71	0,4	0	NC	2,74	**No testeado**	Priorizable
E2F4_HUMAN	Transcription factor E2F4	E**ELMSSEVF**APLLRLS	2B	0,66	0,46	0	N	3,02	W (E2F TP)	Priorizable
E2F4_HUMAN	Transcription factor E2F4	SELLE**ELMSSEVF**APL	2B	0,63	0,46	0	N	3,02	No testeado	Priorizable
PLAL1_HUMAN	Zinc finger protein PLAGL1	VCAL**ELGSTEVL**LDHL	2B	0,45	0,2	0	NC, SE	3,14	No testeado	Priorizable (IUPred)

Uniprot ID	Proteína	Secuencia[1]	Variante[#]	RSA[##]	IUPred[##]	Pfam* [##]	Localización Celular**	FoldX***	Testeado [†]	Criterio aplicado
AKA12_HUMAN	A-kinase anchor protein 12	QVEAEAA**LLTEEVL**ER	3	0,80	0,71	0	C	0,56	No testeado	Priorizable (Localización)
AKA12_HUMAN	A-kinase anchor protein 12	EAA**LLTEEVL**EREVIA	3	0,77	0,73	0	C	0,56	No testeado	Priorizable (Localización)
HDAC7_HUMAN	Histone deacetylase 7	QS**LMTTERL**SGSGLHW	3	0,86	0,58	0	NC	0,95	No testeado	Priorizable
CLMN_HUMAN	Calmin	**LMTVEAL**EEGDYFEAI	3	0,84	0,46	0	C	0,99	No testeado	Priorizable (Localización)
CAC1C_HUMAN	Voltage-dependent L-type calcium channel subunit alpha-1C	SEPS**LLSTEML**SYQDD	3	0,82	0,56	0	C	1,11	No testeado	Priorizable (Localización)
CAC1C_HUMAN	Voltage-dependent L-type calcium channel subunit alpha-1C	**LLSTEML**SYQDDENRQ	3	0,80	0,60	0	C	1,11	No testeado	Priorizable (Localización)
GTSE1_HUMAN	G2 and S phase-expressed protein 1	DIL**LLADEKF**DFDLSL	3	0,75	0,33	15	NC	1,50	SP	Priorizable (IUpred, Pfam)
NCK5L_HUMAN	Nck-associated protein 5-like	G**LETSESL**SDSLYDSL	3	0,86	0,78	0	C	1,89	No testeado	Priorizable (Localización)
S31D1_HUMAN	Spermatogenesis-associated protein 31D1	QQ**LLSWESL**KDAAPSV	3	0,73	0,41	16	C	2,13	N	Priorizable (Pfam, Localización)
E2F2_HUMAN	Transcription factor E2F2	DYLWG**LEAGEGI**SDLF	3	0,65	0,42	0	N	2,13	SP (E2F TP)	Priorizable
BTBDI_HUMAN	BTB/POZ domain-containing protein 18	TGLEVS**LTTDELL**YPS	3	0,87	0,53	0	N	2,16	No testeado	Priorizable
BTBDI_HUMAN	BTB/POZ domain-containing protein 18	VS**LTTDELL**YPSPKAG	3	0,86	0,57	0	N	2,16	No testeado	Priorizable
SDS3_HUMAN	Sin3 histone deacetylase corepressor complex SDS3	NY**LLTDEQI**MEDLRTL	3	0,56	0,44	16	NC	2,32	SP	Priorizable (Pfam)
XPO4_HUMAN	Exportin-4	TN**LLSKEFI**DFSDTDE	3	0,43	0,09	0	NC, SE	2,51	W	Priorizable (IUPred)
ZN436_HUMAN	Zinc finger protein 436	**LTAEEWV**SYPLQPVTD	3	0,63	0,37	0	NC	2,74	No testeado	Priorizable (IUPred)
GATA6_HUMAN	Transcription factor GATA-6	GN**LSSWEDL**LLFTDLD	3	0,87	0,39	0	NC, SE	3,12	SP	Priorizable (IUpred)
GATA6_HUMAN	Transcription factor GATA-6	AGPGGN**LSSWEDL**LLF	3	0,86	0,48	0	NC, SE	3,12	No testeado	Priorizable

Uniprot ID	Proteína	Secuencia[1]	Variante[#]	RSA[##]	IUPred[##]	Pfam* [##]	Localización Celular**	FoldX***	Testeado [†]	Criterio aplicado
LAR1B_HUMAN	La-related protein 1B	LN**LISKEQF**ENLTPEL	3	0,71	0,49	0	N	3,63	No testeado	Priorizable
CE162_HUMAN	Centrosomal protein of 162 kDa	SLLSTDS**LETNELV**VS	3	0,84	0,48	16	NC	4,45	W	Priorizable (Pfam)
UIMC1_HUMAN	BRCA1-A complex subunit RAP80	VCPETQ**LSSSETF**DLE	3	0,89	0,6	0	N	4,54	No testeado	Priorizable
TSC2_HUMAN	Tuberin	TSW**LMSLENPL**SPFSS	3	0,57	0,47	0	NC, SE	4,83	No testeado	Priorizable
ZN445_HUMAN	Zinc finger protein 445	W**LEAREPW**GLNMQAAQ	3	0,62	0,51	0	N	10,05	W	Priorizable (FoldX)
SEPT7_HUMAN	Septin-7	SLFLT**DLYSPE**YPGPS	4A	0,46	0,26	16	NC	4,27	N	Priorizable (IUPred, Pfam)
CE295_HUMAN	Centrosomal protein of 295 kDa	S**LLSYE**NTDLSLTDPE	4B	0,76	0,51	0	C	3,38	SP	Priorizable (Localización)
TARA_HUMAN	TRIO and F-actin-binding protein	W**LLAEE**TAATASAIEA	5	0,52	0,51	0	NC	2,42	No testeado	Priorizable
GASP1_HUMAN	G-protein coupled receptor-associated sorting protein 1	IGSW**LWATEESN**IDGT	5	0,83	0,44	0	C	3,23	No testeado	Priorizable (Localización)
GASP1_HUMAN	G-protein coupled receptor-associated sorting protein 1	PEAIIGSW**LWATE**ESN	5	0,81	0,37	0	C	3,23	No testeado	Priorizable (IUPred, Localización)
IKKB_HUMAN	Inhibitor of nuclear factor kappa-B kinase subunit beta	ALDWSW**LQTEE**EEHSC	5	0,77	0,43	7	NC	5,09	No testeado	Priorizable (FoldX)
SGSM1_HUMAN	Small G protein signaling modulator 1	**DLLANESM**DEFMSITG	2B	0,92	0,58	0	SE	-0,97	No testeado	No priorizable (Localización)
TLXNB_HUMAN	Putative TLX1 neighbor protein	HS**LLSQEAM**GPGEGAE	3	0,85	0,77	16	O	0,37	W	No priorizable (Localización)
DC8L2_HUMAN	DDB1- and CUL4-associated factor 8-like protein 2	HF**LMSGESL**FHYPLVG	3	0,52	0,70	0	O	1,32	W	No priorizable (Localización)
MACF1_HUMAN	Microtubule-actin cross-linking factor 1, isoforms 1/2/3/5	QW**LESKEEV**LKSMDAM	3	0,59	0,48	16	SE	4,7	No testeado	No priorizable (Localización)
CU024_HUMAN	Putative uncharacterized protein encoded by LINC00114	GCSYPTSW**LSSQESF**S	3	0,44	0,37	0	O	4,87	No testeado	No priorizable (Localización)

Uniprot ID	Proteína	Secuencia[1]	Variante[#]	RSA[##]	IUPred[##]	Pfam* [##]	Localización Celular**	FoldX***	Testeado [†]	Criterio aplicado
TM1L1_HUMAN	TOM1-like protein 1	MN**LLALE**NTEIPPFAQ	5	0,85	0,61	0	SE	2,04	No testeado	No priorizable (Localización)
ABCA8_HUMAN	ATP-binding cassette sub-family A member 8	SHLLFSS**LFSEERM**DV	2A	0,56	0,01	16	C	0,28	No testeado	Filtrado (IUPred, Pfam)
NCKX1_HUMAN	Sodium/potassium/calcium exchanger 1	S**LFSREIL**NLTWWPLF	2A	0,31	0,01	16	C	1,59	No testeado	Filtrado (RSA, IUPred, Pfam)
XPO1_HUMAN	Exportin-1	**LLSEEV**FDFSSGQITQ	3	0,22	0,07	16	NC, SE	2,78	No testeado	Filtrado (RSA, IUPred, Pfam)
UTRN_HUMAN	Utrophin	EAASLSEW**LSATETEL**	3	0,33	0,38	16	NC	4,79	No testeado	Filtrado (RSA)
DMD_HUMAN	Dystrophin	QW**LEAKE**EAEQVLGQA	5	0,3	0,43	16	NC	3,75	No testeado	Filtrado (RSA)

[1]Se destaca en negrita la variante de expresión regular y en subrayado la subsecuencia a la cual corresponde el valor de FoldX de cada péptido

[#]Variante 1: `[DE][IL][YFH][CAST].E.{1,2}[WFILYVM]` o `[DE][IL].[CAST][YFH]E.{1,2}[WFILYVM];`

Variante 2A: `[IL][YFH][CAST].E.{1,2}[WFILYVM]` o `[IL].[CAST][YFH]E.{1,2}[W FILM];`

Variante 2B: `[DE][IL].[CAST].E.{1,2}[WFILYVM];`

Variante 2C: `[DE][IL][YFH][CAST].E` o `[DE][IL].[CAST][YFH]E`

Variante 3: `[IL].[CAST].E.{1,2}[WFILYVM]`

Variante 4A: `[DE][IL].[CAST].E`

Variante 4B: `[IL][YFH][CAST].E.` o `[IL].[CAST][YFH]E.`

Variante 5: `[IL].[CAST].E`

[##]Celdas en amarillo indican una advertencia. Celdas en rojo señalan el motivo de filtrado.

* Número de residuos solapados con dominios Pfam

** N= Núcleo; NC= Núcleo y Citoplasma, C= Citoplasma; SE= Sistema Endomembrana; O= Otro

***Valor mínimo de secuencia escaneada con matriz FoldX 1GUX_8.

[†]SP= Péptido de interacción *Strong Positive* de acuerdo a ensayos del laboratorio, W = Péptido de interacción *Weak* de acuerdo a ensayos del laboratorio, N= Péptido sin interacción de acuerdo a ensayos del laboratorio, E2F TP = Péptido conteniendo al SLiM E2F reportado en ELM como verdadero positivo, LxCxE TP = Péptido conteniendo al SLiM LxCxE reportado en ELM como verdadero positivo

[%]El péptido de KDM5A_HUMAN presenta dos Variantes de regex (2A, siete posiciones en negrita correspondiente al SLiM funcional y 5, cinco posiciones subrayadas). El mínimo valor de FoldX corresponde a la Variante 5, donde la matriz incluye sólo las 5 posiciones subrayadas. La variante 2A corresponde al siguiente mínimo de FoldX (Subsecuencia: `LFCDEEIP` ; FoldX= 4,52)

Tabla S24. Criterios de priorización aplicados a péptidos *hit* de p107 con patrones de secuencia del SLiM E2F.

Uniprot ID	Proteína	Secuencia[1]	Variante[#]	RSA[##]	IUPred[##]	Pfam* [##]	Localización Celular**	FoldX***	Testeado [†]	Criterio aplicado
E2F2_HUMAN	Transcription factor E2F2	DYLWGLEAGEG**ISDLF**	1	0,65	0,42	0	N	0	SP (E2F TP)	Priorizable
KIF15_HUMAN	Kinesin-like protein KIF15	STQ**MQELF**SSERIDWT	1	0,57	0,46	0	C	0,56	**No testeado**	Priorizable (Localización)
MYT1L_HUMAN	Myelin transcription factor 1-like protein	YVTT**LTEMY**TNQDRYQ	1	0,41	0,54	0	N	1,18	SP	Priorizable
SEPT7_HUMAN	Septin-7	SLF**LTDLY**SPEYPGPS	1	0,46	0,26	16	NC	1,7	N	Priorizable (IUPred, Pfam)
E2F3_HUMAN	Transcription factor E2F3	SD**LFDAY**DLEKLPLVE	1	0,73	0,31	0	NC	5,99	SP (E2F TP)	Priorizable (IUPred, FoldX)
E2F4_HUMAN	Transcription factor E2F4	SEL**LEELM**SSEVFAPL	3	0,63	0,46	0	N	4,42	SP (E2F TP)	Priorizable (FoldX)
ZN436_HUMAN	Zinc finger protein 436	QWGDLT**AEEWV**SYPLQ	3	0,71	0,4	0	NC	5,6	**No testeado**	Priorizable (FoldX)
ZN436_HUMAN	Zinc finger protein 436	LT**AEEWV**SYPLQPVTD	3	0,63	0,37	0	NC	5,6	No testeado	Priorizable (IUPred, FoldX)
NEBU_HUMAN	Nebulin	KH**AMEVA**KKQSDVAYR	3	0,55	0,57	0	C	5,98	**No testeado**	Priorizable (Localización, FoldX)
S31D1_HUMAN	Spermatogenesis-associated protein 31D1	QQLLSWES**LKDAA**PSV	3	0,73	0,41	16	C	8,41	N	Priorizable (Pfam, Localización, FoldX)
INKA1_HUMAN	PAK4-inhibitor INKA1	LVLGDNCFADL**VHNWM**	4	0,54	0,32	16	NC	1,65	**No testeado**	Priorizable (IUPred, Pfam)
SGSM1_HUMAN	Small G protein signaling modulator 1	DLLANES**MDEFM**SITG	3	0,92	0,58	0	SE	1,09	No testeado	No priorizable (Localización)
TM9S4_HUMAN	Transmembrane 9 superfamily member 4	MAT**AMDWL**PWSLLLFS	3	0,68	0,02	0	SE	2,06	No testeado	No priorizable (Localización)
LAP4A_HUMAN	Lysosomal-associated transmembrane protein 4A	EV**IGNYY**SSERMADNA	2	0,55	0,02	16	SE	-0,55	No testeado	Filtrado (IUPred, Pfam)
OR5T1_HUMAN	Olfactory receptor 5T1	VPI**IGDFW**LHSPMYYF	3	0,3	0	16	SE	2,13	No testeado	Filtrado (RSA, IUPred, Pfam)
UTRN_HUMAN	Utrophin	EAAS**LSEWL**SATETEL	3	0,33	0,38	16	NC	4,25	No testeado	Filtrado (RSA)

Uniprot ID	Proteína	Secuencia[1]	Variante[#]	RSA[##]	IUPred[##]	Pfam* [##]	Localización Celular**	FoldX***	Testeado [†]	Criterio aplicado
ZNG1C_HUMAN	COBW domain-containing protein 3	**LSNVL**DLHAFDSLSGI	4	0,23	0,14	0	NC, C, O	4,6	No testeado	Filtrado (RSA)
DMD_HUMAN	Dystrophin	QWLEAKEE**AEQVL**GQA	4	0,3	0,43	16	NC	6,41	No testeado	Filtrado (RSA)

[1]Se destaca en negrita la variante de expresión regular y en subrayado la subsecuencia a la cual corresponde el valor de FoldX de cada péptido.

[#] Variante 1: `[IVLMA].[DE][IVLMAFYW][FY]`

Variante 2: `[IVLMA].[NQDE][IVLMAFYW][FY]`

Variante 3: `[IVLMA].[DE][IVLMAFYW][IVLMAFYW]`

Variante 4: `[IVLMA].[NQDE][IVLMAFYW][IVLMAFYW]`

[##]Celdas en amarillo indican una advertencia. Celdas en rojo señalan el motivo de filtrado.

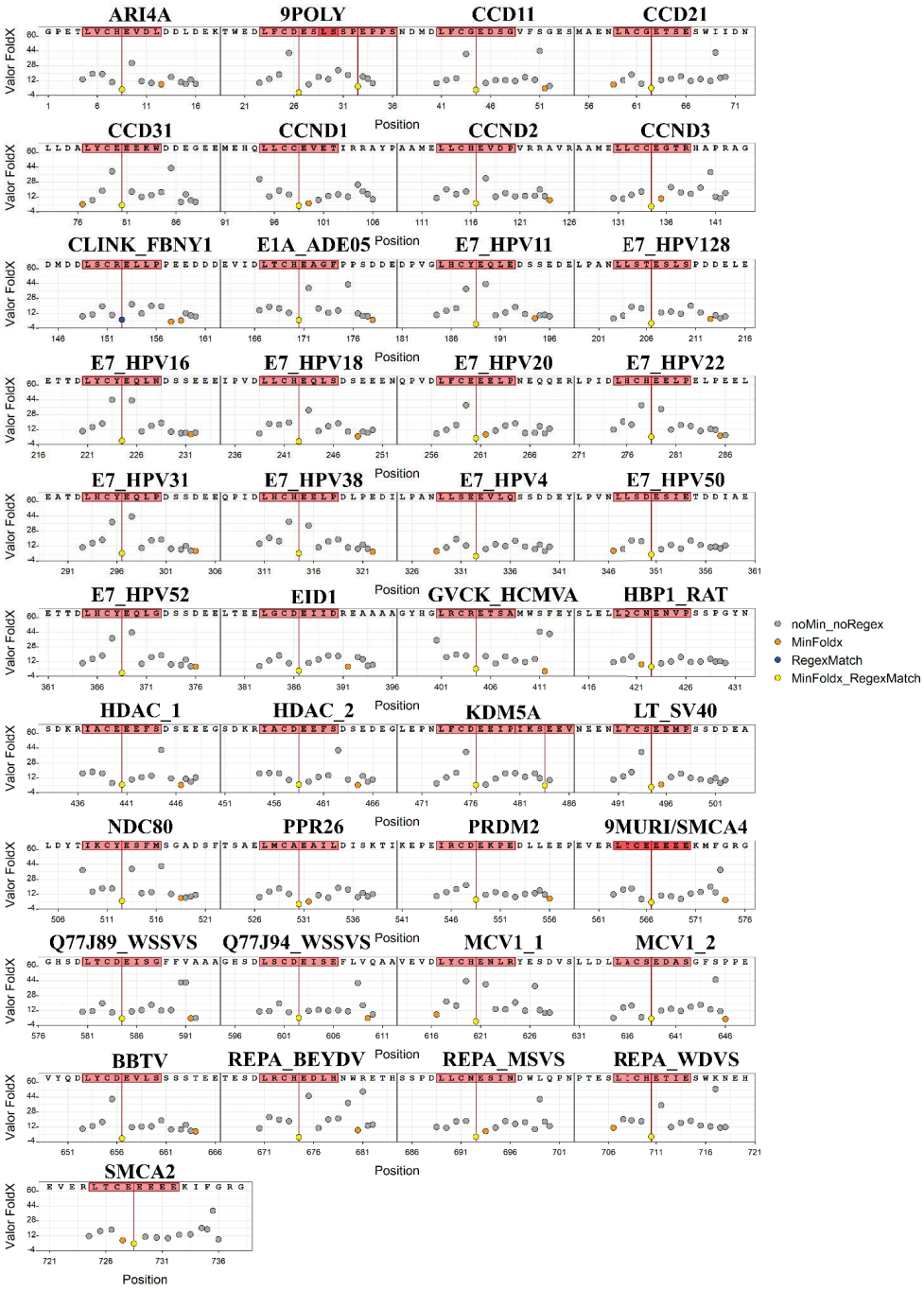
* Número de residuos solapados con dominios Pfam

** N= Núcleo; NC= Núcleo y Citoplasma, C= Citoplasma; SE= Sistema Endomembrana; O= Otro

***Valor mínimo de secuencia escaneada con matriz FoldX 1N4M_5.

[†]SP= Péptido de interacción *Strong Positive* de acuerdo a ensayos del laboratorio, W = Péptido de interacción *Weak* de acuerdo a ensayos del laboratorio, N= Péptido sin interacción de acuerdo a ensayos del laboratorio, E2F TP = Péptido conteniendo al SLiM E2F reportado en ELM como verdadero positivo

A


ARI4A
9POLY
CCD11
CCD21
CCD31
CCND1
CCND2
CCND3
CLINK_FBNY1
E1A_ADE05
E7_HPV11
E7_HPV128
E7_HPV16
E7_HPV18
E7_HPV20
E7_HPV22
E7_HPV31
E7_HPV38
E7_HPV4
E7_HPV50
E7_HPV52
EID1
GVCK_HCMVA
HBP1_RAT
HDAC_1
HDAC_2
KDM5A
LT_SV40
NDC80
PPR26
PRDM2
9MURI/SMCA4
Q77J89_WSSVS
Q77J94_WSSVS
MCV1_1
MCV1_2
BBTV
REPA_BEYDV
REPA_MSVS
REPA_WDVS
SMCA2
Valor FoldX
Position
noMin_noRegex
MinFoldx
RegexMatch
MinFoldx_RegexMatch

B

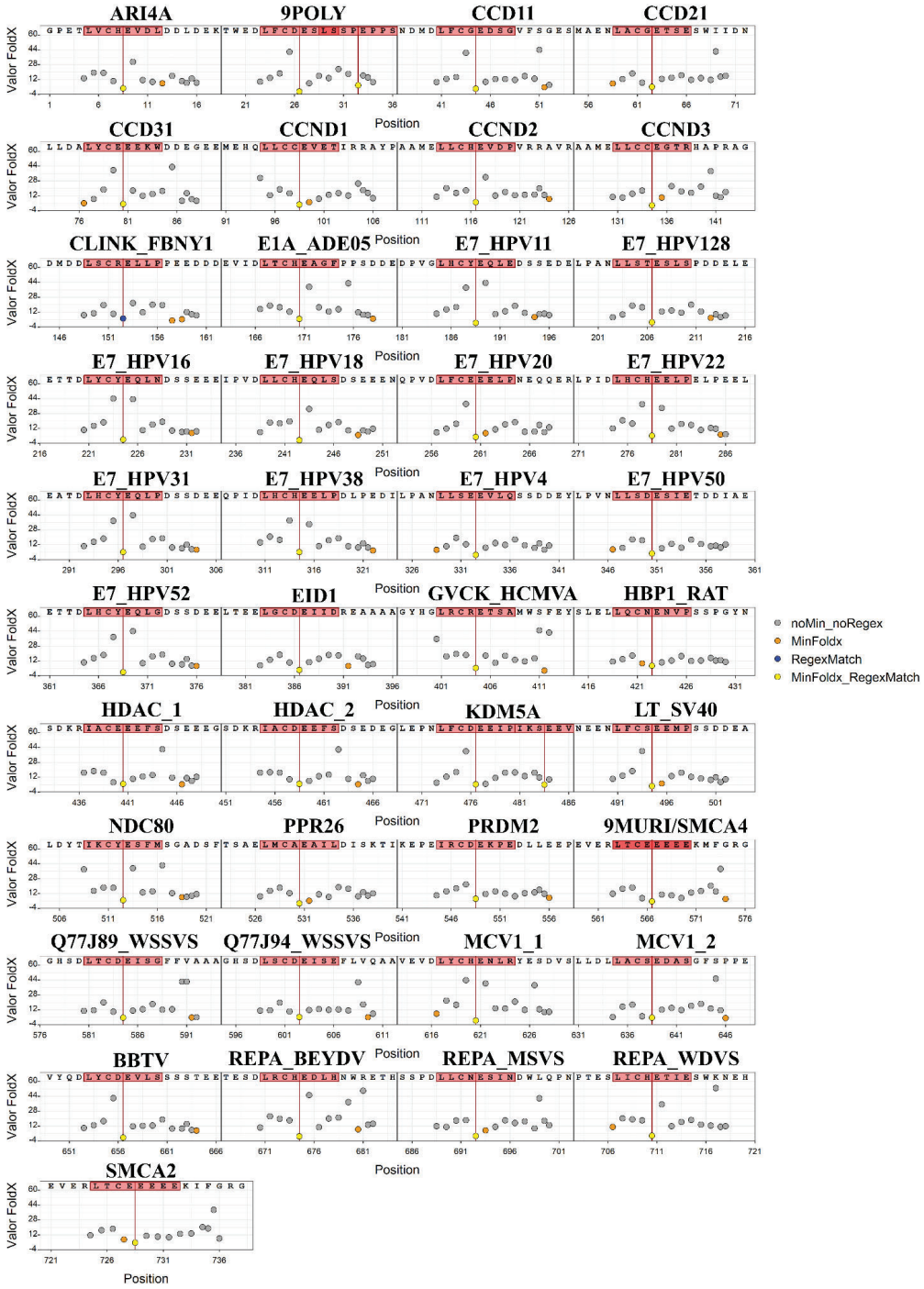
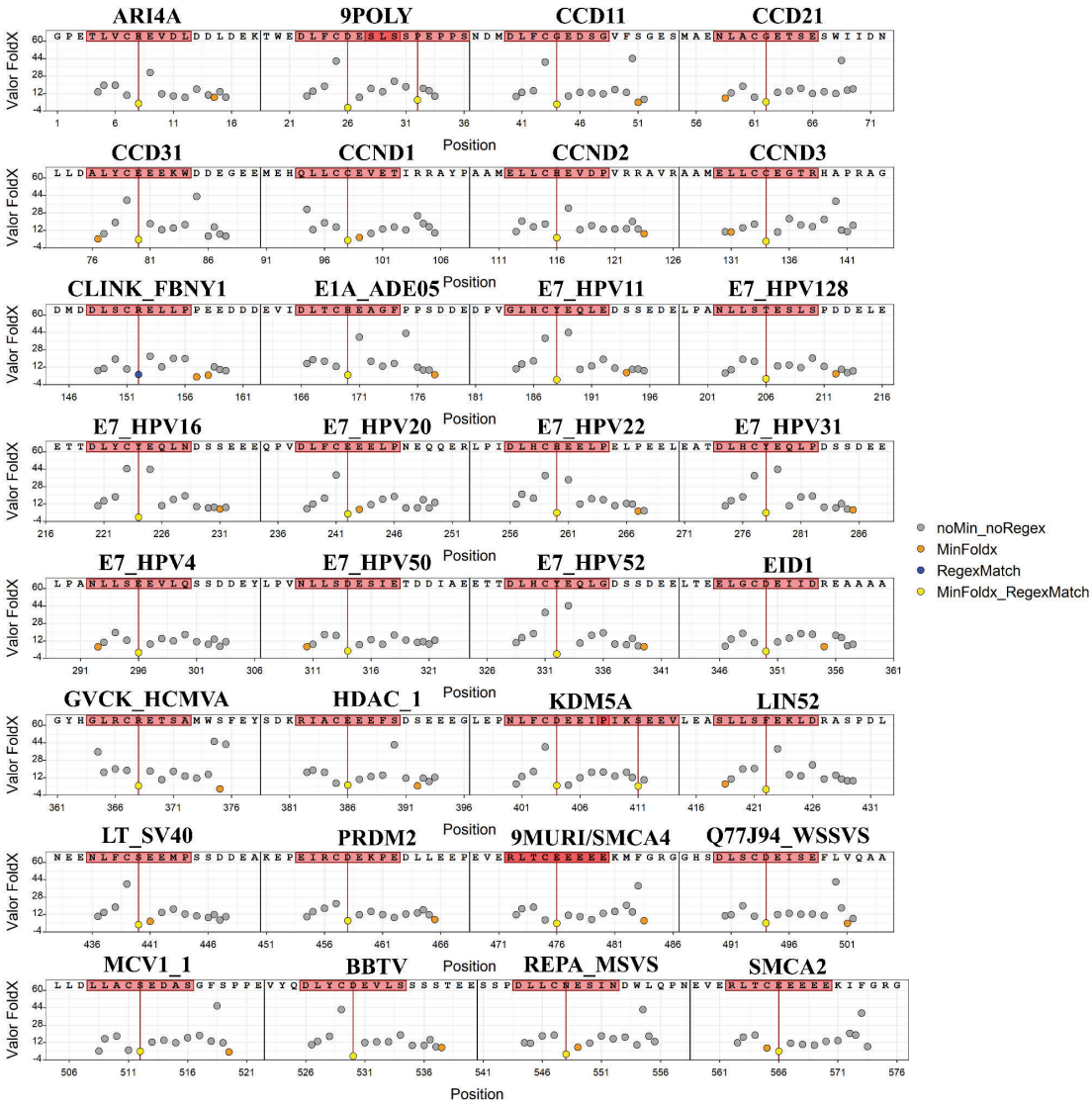


Figura S1. Distribución de valores FoldX de péptidos TP con SLiM LxCxE reportados para Rb escaneados con variantes de 1GUX. Gráfico de puntos de péptidos conocidos de interacción con la proteína *pocket* Rb. Se indica sobre el recuadro el nombre de las 42 proteínas cuyos péptidos de 18 residuos fueron escaneados con **A:** 1GUX_9, **B:** 1GUX_8 y se encuentran reportadas como TP en la base de datos ELM [36]. El eje *x* indica la posición como el punto medio entre la posición inicial y final de cada sub-secuencia analizada (abajo) y el detalle de residuo de la secuencia peptídica (arriba). Las secuencias se organizaron de manera continua una de otras y se marcó el inicio de una secuencia distinta con una línea vertical negra. Las líneas verticales rojas indican el punto medio de la subsecuencia donde se detectó la expresión regular `[IL].[CAST].E`. Los puntos amarillos indican la detección de una subsecuencia conteniendo al SLiM LxCxE (recuadro rojo), con el valor mínimo de FoldX de todas las subsecuencias escaneadas para ese péptido.

A


ARI4A
9POLY
CCD11
CCD21
CCD31
CCND1
CCND2
CCND3
CLINK_FBNY1
E1A_ADE05
E7_HPV11
E7_HPV128
E7_HPV16
E7_HPV20
E7_HPV22
E7_HPV31
E7_HPV4
E7_HPV50
E7_HPV52
EID1
GVCK_HCMVA
HDAC_1
KDM5A
LIN52
LT_SV40
PRDM2
9MURI/SMCA4
Q77J94_WSSVS
MCV1_1
BBTV
REPA_MSVS
SMCA2
Valor FoldX
Position
noMin_noRegex
MinFoldx
RegexMatch
MinFoldx_RegexMatch

B

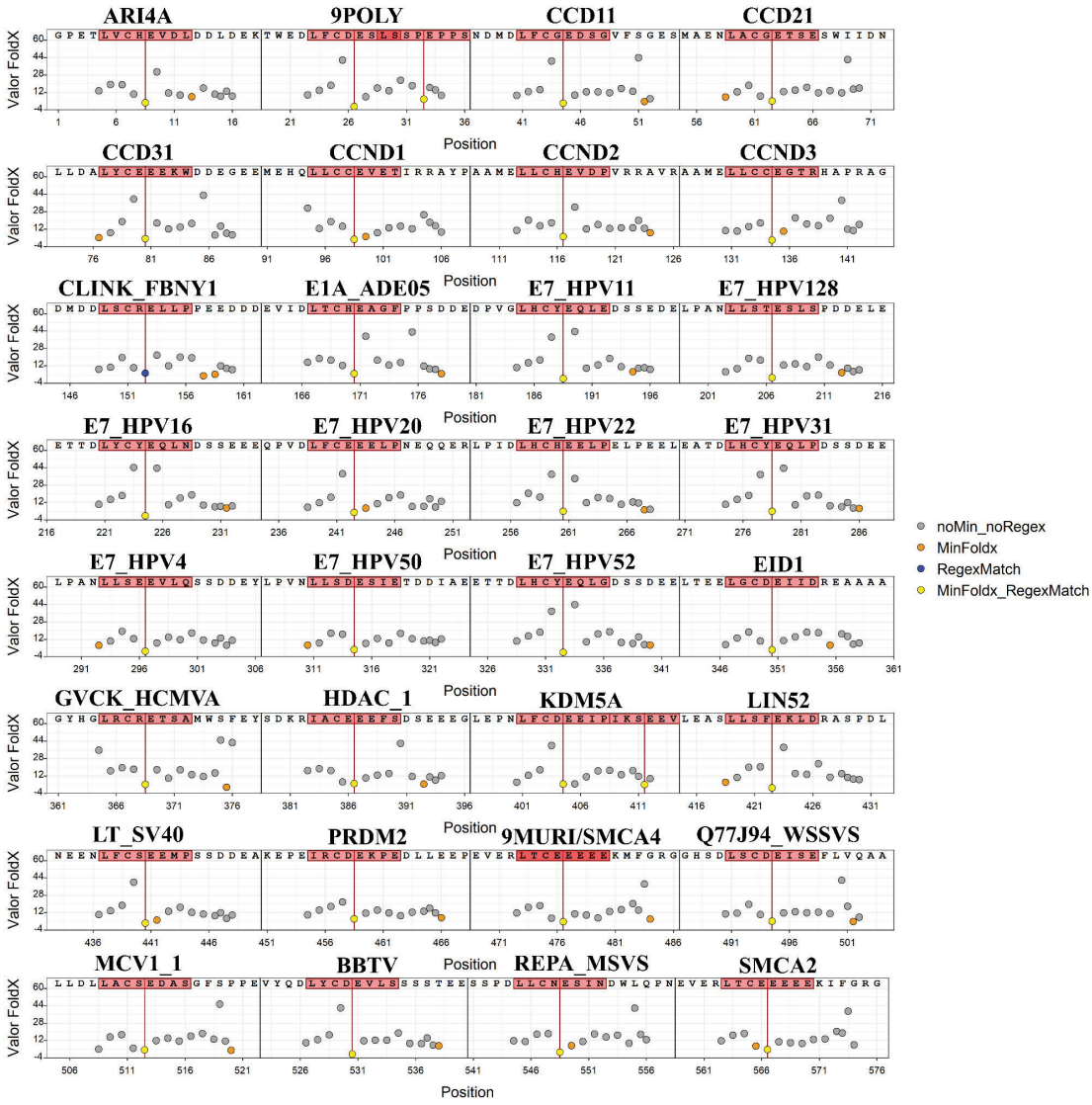


Figura S2. Distribución de valores FoldX de péptidos TP con SLiM LxCxE reportados para p107 escaneados con variantes de 1GUX. Gráfico de puntos de péptidos conocidos de interacción con la proteína *pocket* p107. Se indica sobre el recuadro el nombre de las 33 proteínas cuyos péptidos de 18 residuos fueron escaneados con **A:** 1GUX_9, **B:** 1GUX_8 y se encuentran reportadas como TP en la base de datos ELM [36] . El eje *x* indica la posición como el punto medio entre la posición inicial y final de cada sub-secuencia analizada (abajo) y el detalle de residuo de la secuencia peptídica (arriba). Las secuencias se organizaron de manera continua una de otras y se marcó el inicio de una secuencia distinta con una línea vertical negra. Las líneas verticales rojas indican el punto medio de la subsecuencia donde se detectó la expresión regular `[IL].[CAST].E`. Los puntos amarillos indican la detección de una subsecuencia conteniendo al SLiM LxCxE (recuadro rojo), con el valor mínimo de FoldX de todas las subsecuencias escaneadas para ese péptido.

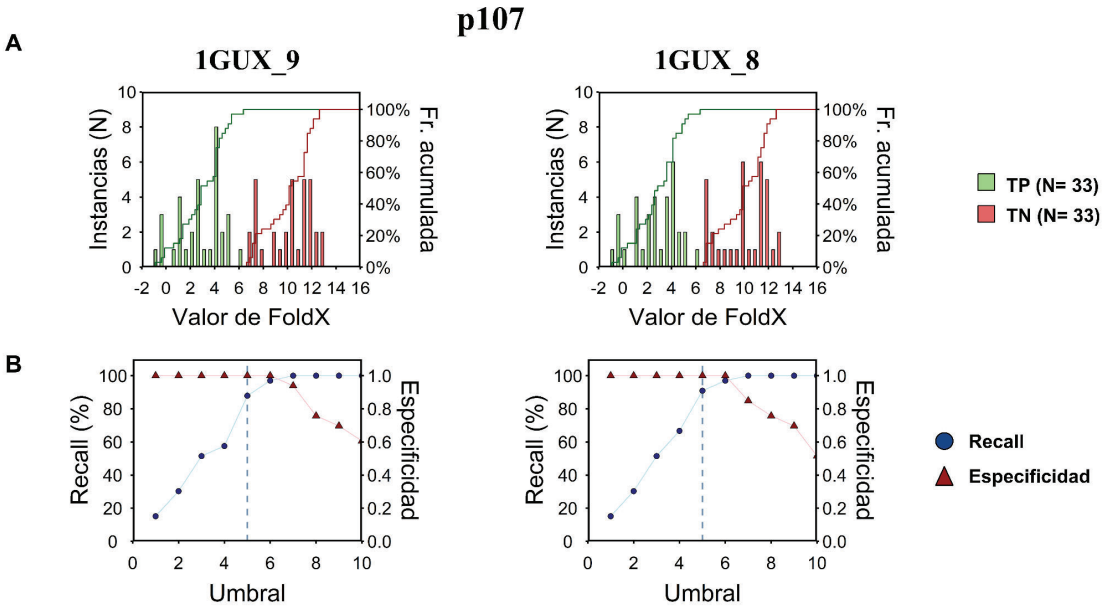


Figura S3. Interactores conocidos de p107 escaneados con variantes 1GUX. A: Distribución de valores FoldX de péptidos TP y TN para matrices 1GUX_9 (izquierda) y 1GUX_8 (derecha). Eje Y izquierdo: Número de instancias, Eje Y derecho: porcentaje de frecuencia acumulada de instancias. B: Recall (puntos azules) y especificidad (triángulos en rojo) de matrices 1GUX_9 (izquierda) y 1GUX_8 (derecha) para diferentes valores umbral de FoldX. La línea vertical punteada señala el valor de umbral utilizado para establecer comparaciones.

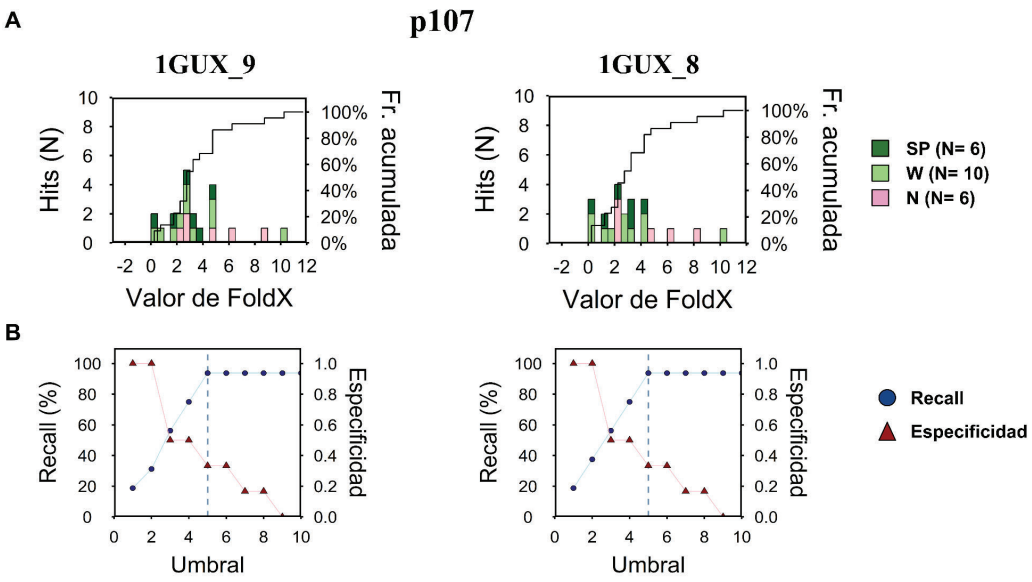


Figura S4. Péptidos *hit* en ProP-PD usando p107 como carnada que fueron testeados experimentalmente y escaneados con variantes 1GUX. A: Distribución de valores FoldX de péptidos TP y TN para matrices 1GUX_9 (izquierda) y 1GUX_8 (derecha). Eje Y izquierdo: Número de instancias, Eje Y derecho: porcentaje de frecuencia acumulada de instancias. **B:** R*ecall* (puntos azules) y especificidad (triángulos en rojo) de matrices 1GUX_9 (izquierda) y 1GUX_8 (derecha) para diferentes valores umbral de FoldX. La línea vertical punteada señala el valor de umbral utilizado para establecer comparaciones.

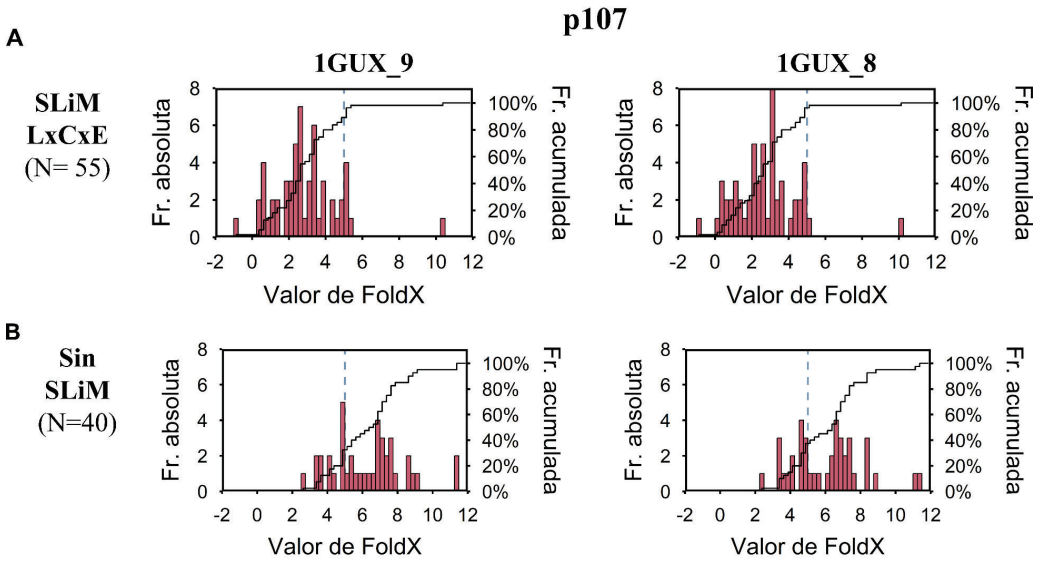
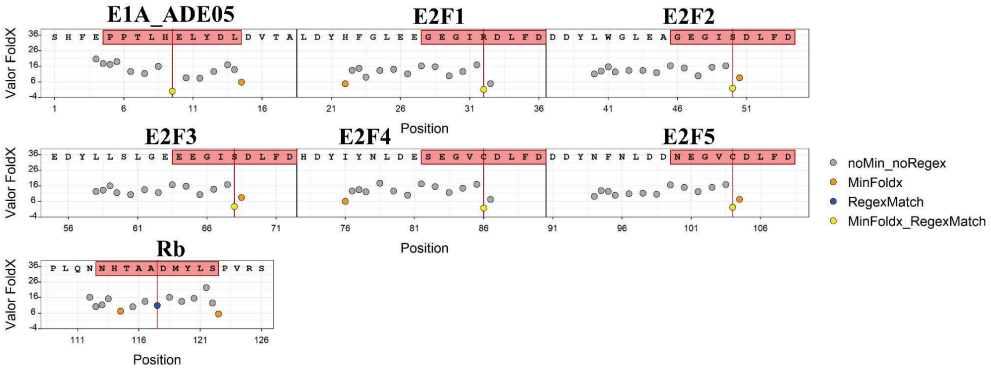
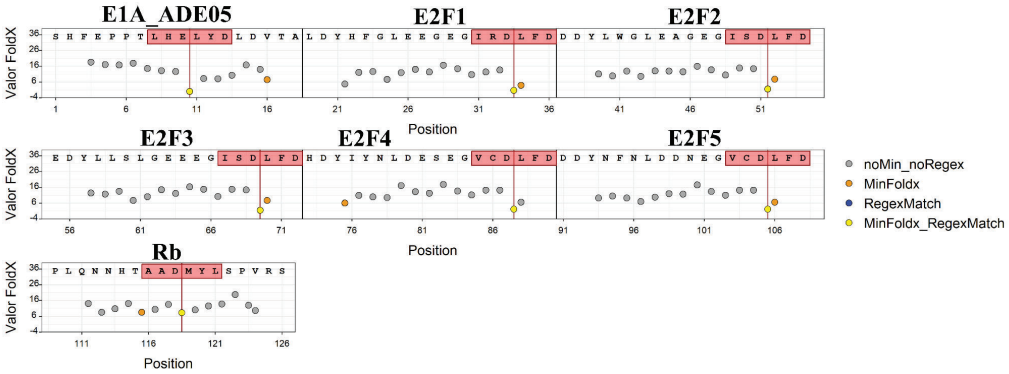


Figura S5. Péptidos *hit* de p107 escaneados con variantes 1GUX. A: Distribución de valores FoldX de péptidos para matrices 1GUX_9 (izquierda) y 1GUX_8 (derecha). Eje Y izquierdo: Frecuencia absoluta de instancias, Eje Y derecho: porcentaje de frecuencia acumulada de instancias. La línea vertical punteada señala el valor de umbral utilizado para establecer comparaciones. **A:** Péptidos con SLiM LxCxE detectado **B:** Péptidos sin SLiM LxCxE detectado.

A



B



C

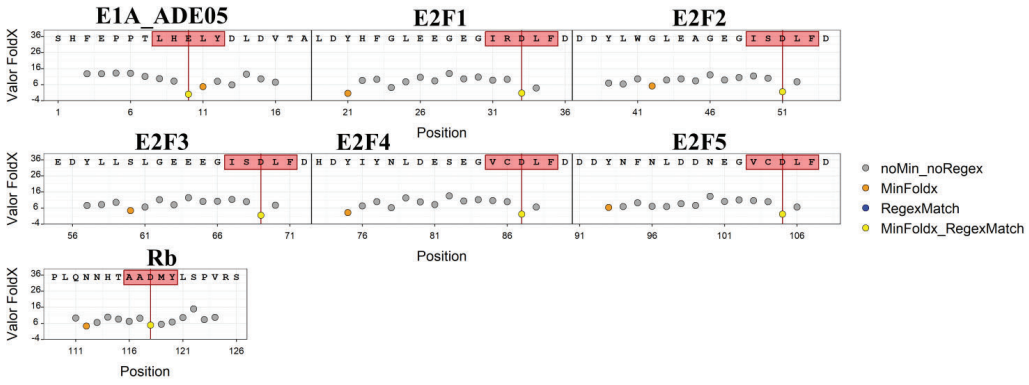
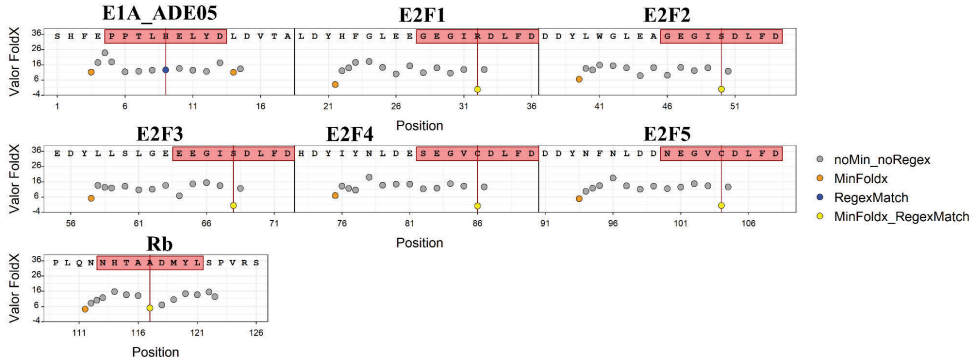
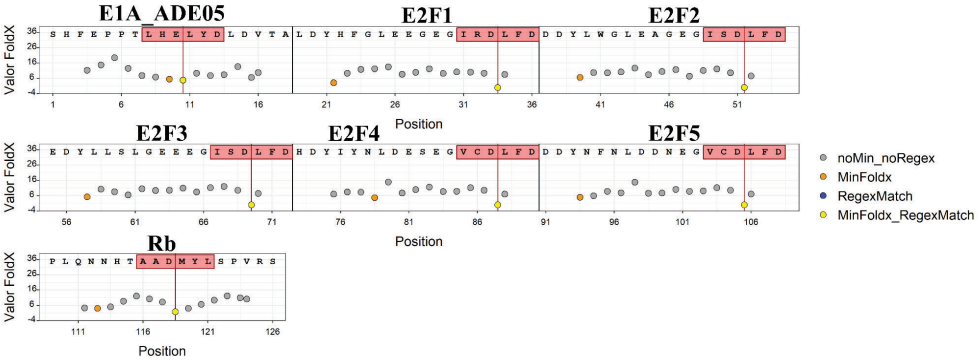


Figura S6. Distribución de valores FoldX de péptidos TP con SLiM E2F reportados para Rb escaneados con variantes de 2R7G. Gráfico de puntos de péptidos conocidos de interacción con la proteína *pocket* Rb. Se indica sobre el recuadro el nombre de las siete proteínas cuyos péptidos de 18 residuos fueron escaneados con **A:** 2R7G_10, **B:** 2R7G_6, **C:** 2R7G_5 y se encuentran reportadas como TP en la base de datos ELM [36]. El eje *x* indica la posición como el punto medio entre la posición inicial y final de cada sub-secuencia analizada (abajo) y el detalle de residuo de la secuencia peptídica (arriba). Las secuencias se organizaron de manera continua una de otras y se marcó el inicio de una secuencia distinta con una línea vertical negra. Las líneas verticales rojas indican el punto medio de la subsecuencia donde se detectó la expresión regular `[IVLMA].[NQDE][IVLFMYAW][IVLFMYAW]`. Los puntos amarillos indican la detección de una subsecuencia conteniendo al SLiM E2F (recuadro rojo), con el valor mínimo de FoldX de todas las subsecuencias escaneadas para ese péptido.

A



B



C

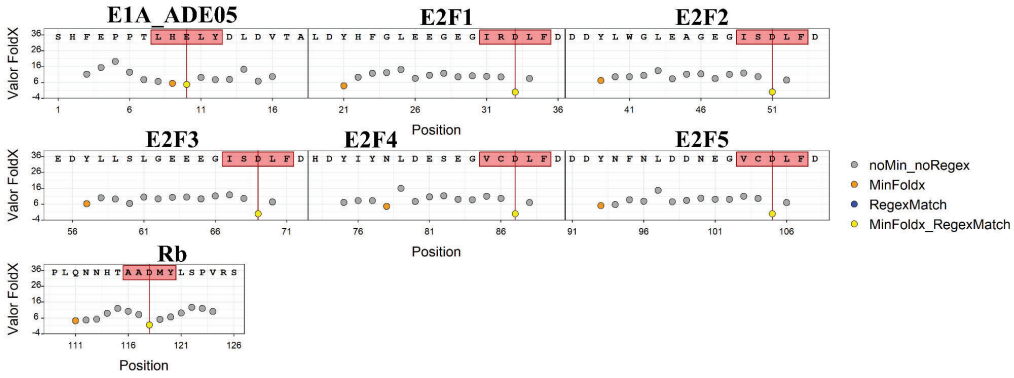
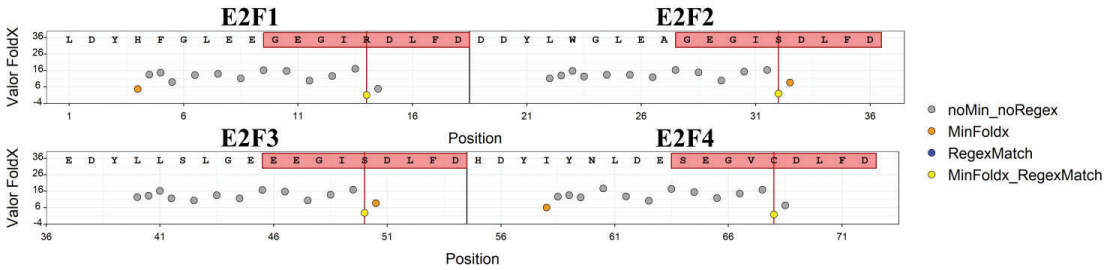
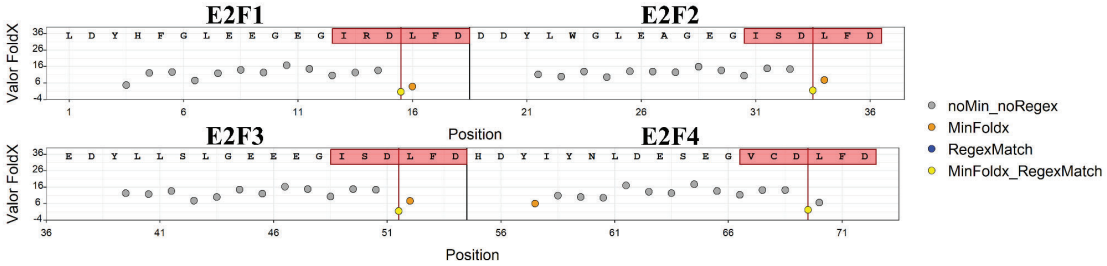


Figura S7. Distribución de valores FoldX de péptidos TP con SLiM E2F reportados para Rb escaneados con variantes de 1N4M. Gráfico de puntos de péptidos conocidos de interacción con la proteína *pocket* Rb. Se indica sobre el recuadro el nombre de las siete proteínas cuyos péptidos de 18 residuos fueron escaneados con **A:** 1N4M_9, **B:** 1N4M_6, **C:** 1N4M_5 y se encuentran reportadas como TP en la base de datos ELM [36] . El eje *x* indica la posición como el punto medio entre la posición inicial y final de cada sub-secuencia analizada (abajo) y el detalle de residuo de la secuencia peptídica (arriba). Las secuencias se organizaron de manera continua una de otras y se marcó el inicio de una secuencia distinta con una línea vertical negra. Las líneas verticales rojas indican el punto medio de la subsecuencia donde se detectó la expresión regular `[IVLMA].[NQDE][IVLFMYAW][IVLFMYAW]`. Los puntos amarillos indican la detección de una subsecuencia conteniendo al SLiM E2F (recuadro rojo), con el valor mínimo de FoldX de todas las subsecuencias escaneadas para ese péptido.

A



B



C

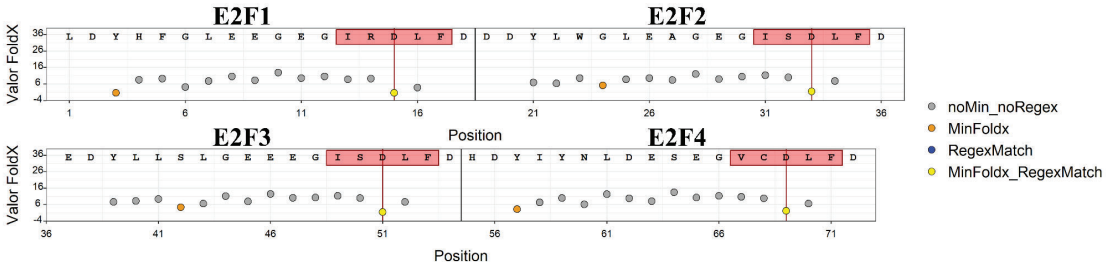
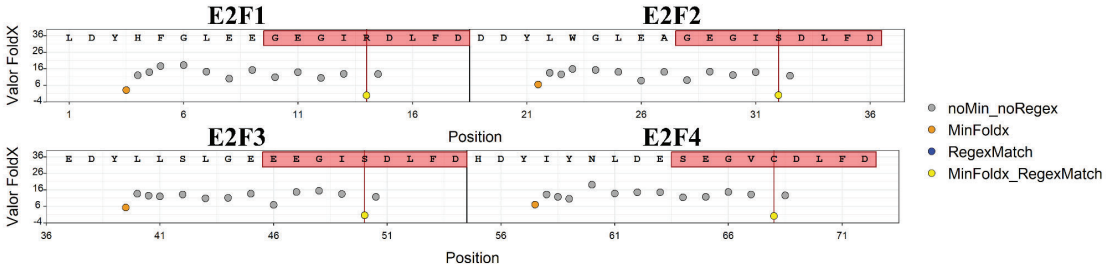
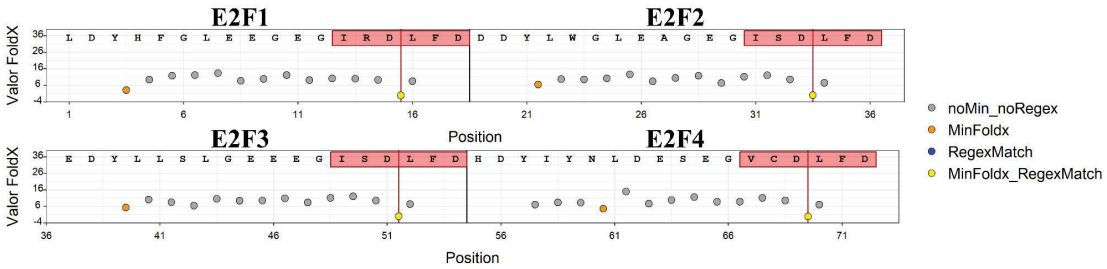


Figura S8. Distribución de valores FoldX de péptidos TP con SLiM E2F reportados para p107 escaneados con variantes de 2R7G. Gráfico de puntos de péptidos conocidos de interacción con la proteína *pocket* p107. Se indica sobre el recuadro el nombre de las cuatro proteínas cuyos péptidos de 18 residuos fueron escaneados con **A:** 2R7G_10, **B:** 2R7G_6, **C:** 2R7G_5 y se encuentran reportadas como TP en la base de datos ELM [36]. El eje *x* indica la posición como el punto medio entre la posición inicial y final de cada sub-secuencia analizada (abajo) y el detalle de residuo de la secuencia peptídica (arriba). Las secuencias se organizaron de manera continua una de otras y se marcó el inicio de una secuencia distinta con una línea vertical negra. Las líneas verticales rojas indican el punto medio de la subsecuencia donde se detectó la expresión regular `[IVLMA].[NQDE][IVLFMYAW][IVLFMYAW]`. Los puntos amarillos indican la detección de una subsecuencia conteniendo al SLiM E2F (recuadro rojo), con el valor mínimo de FoldX de todas las subsecuencias escaneadas para ese péptido.

A



B



C

Figura S9. Distribución de valores FoldX de péptidos TP con SLiM E2F reportados para p107 escaneados con variantes de 1N4M. Gráfico de puntos de péptidos conocidos de interacción con la proteína *pocket* p107. Se indica sobre el recuadro el nombre de las cuatro proteínas cuyos péptidos de 18 residuos fueron escaneados con **A:** 1N4M_9, **B:** 1N4M_6, **C:** 1N4M_5 y se encuentran reportadas como TP en la base de datos ELM [36]. El eje *x* indica la posición como el punto medio entre la posición inicial y final de cada sub-secuencia analizada (abajo) y el detalle de residuo de la secuencia peptídica (arriba). Las secuencias se organizaron de manera continua una de otras y se marcó el inicio de una secuencia distinta con una línea vertical negra. Las líneas verticales rojas indican el punto medio de la subsecuencia donde se detectó la expresión regular `[IVLMA].[NQDE][IVLFMYAW][IVLFMYAW]`. Los puntos amarillos indican la detección de una subsecuencia conteniendo al SLiM E2F (recuadro rojo), con el valor mínimo de FoldX de todas las subsecuencias escaneadas para ese péptido.

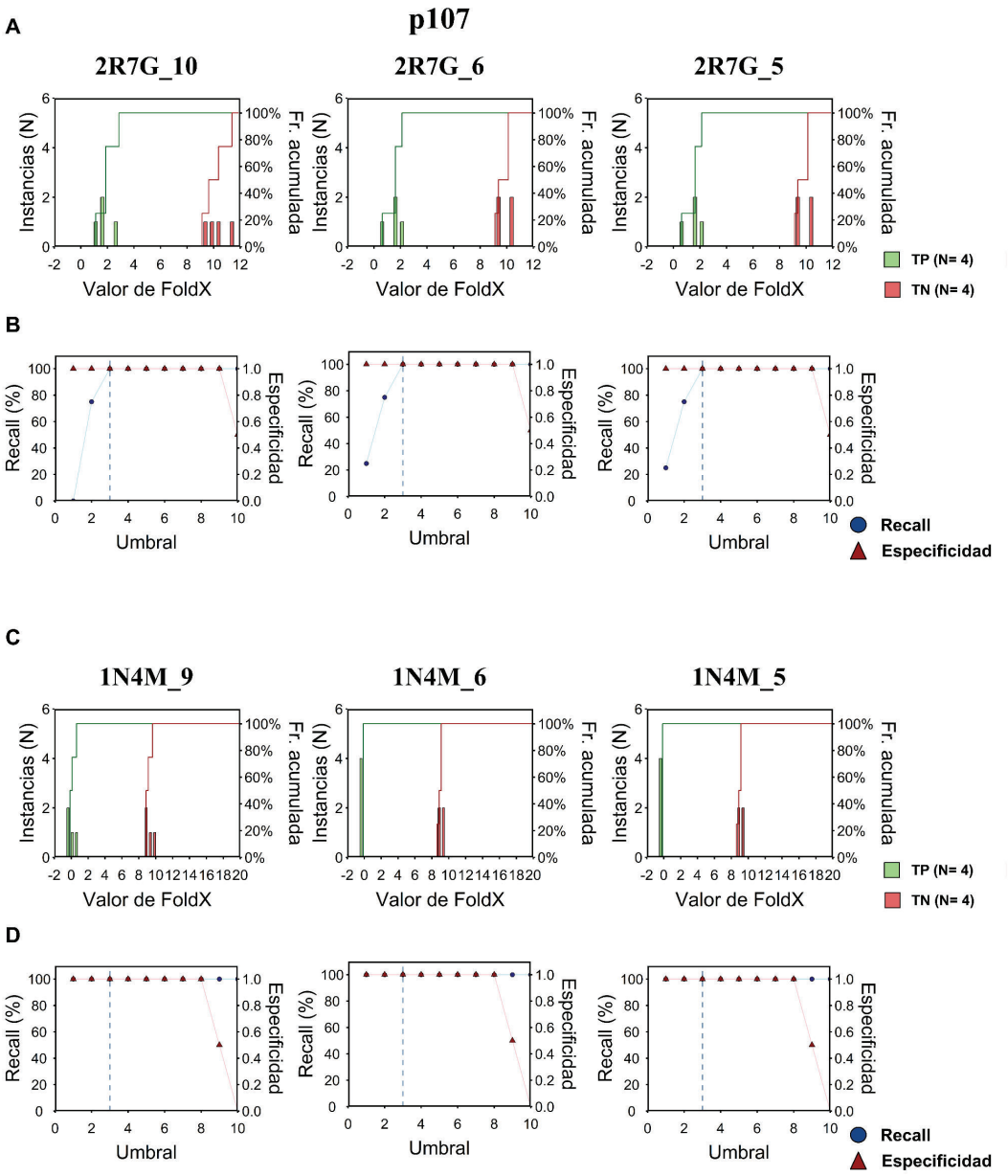


Figura S10. Interactores conocidos TP y TN de p107 escaneados con variantes 2R7G y 1N4M. Distribución de valores FoldX de péptidos TP y TN para matrices **A:** 2R7G y **C:** 1N4M. Eje Y izquierdo: Número de instancias, Eje Y derecho: porcentaje de frecuencia acumulada de instancias. **B:** R*ecall* (puntos azules) y especificidad (triángulos en rojo) de matrices 2R7G_10 (izquierda), 2R7G_6 (medio) y 2R7G_5 (derecha) para diferentes valores umbral de FoldX. **D:** Detalle del *recall* (puntos azules) y especificidad (triángulos rojos) de 1N4M_9 (izquierda), 1N4M_6 (medio) y 1N4M_5 (derecha). En ambos casos la línea vertical punteada señala un valor estimado de umbral de tres para establecer comparaciones.

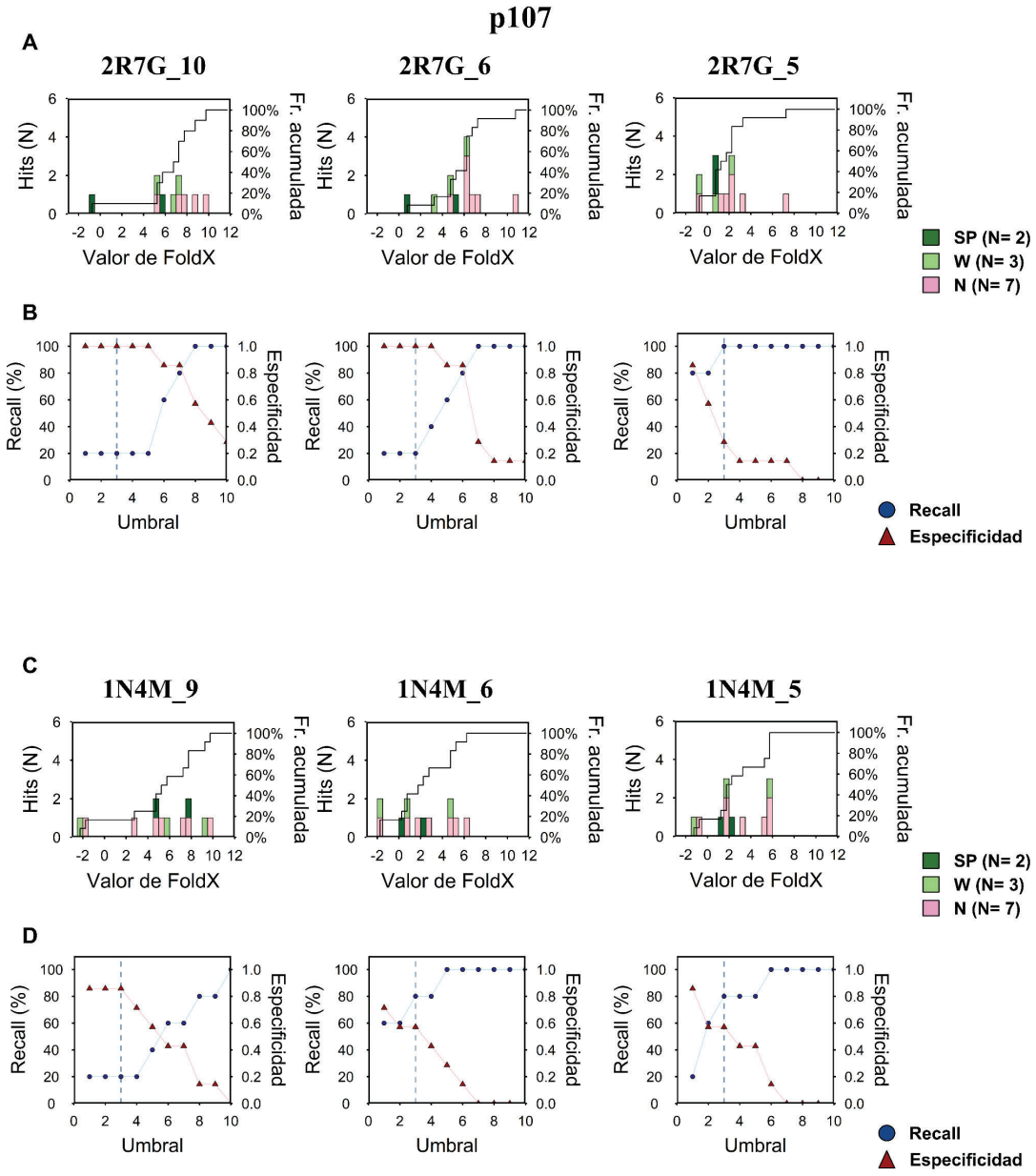


Figura S11. Péptidos *hit* en ProP-PD usando p107 como carnada que fueron testeados experimentalmente y escaneados con variantes de 2R7G y 1N4M. Distribución de péptidos testeados experimentalmente de acuerdo a valores FoldX de las variantes **A:** 2R7G y **C:** 1N4M. Eje Y izquierdo: Número de instancias, Eje Y derecho: porcentaje de frecuencia acumulada de instancias. **B:** R*ecall* (puntos azules) y especificidad (triángulos en rojo) de matrices 2R7G_10 (izquierda), 2R7G_6 (medio) y 2R7G_5 (derecha) para diferentes valores umbral de FoldX. **D:** Detalle del *recall* (puntos azules) y especificidad (triángulos rojos) de 1N4M_9 (izquierda), 1N4M_6 (medio) y 1N4M_5 (derecha). En ambos casos la línea vertical punteada señala un valor estimado de umbral de tres para establecer comparaciones.

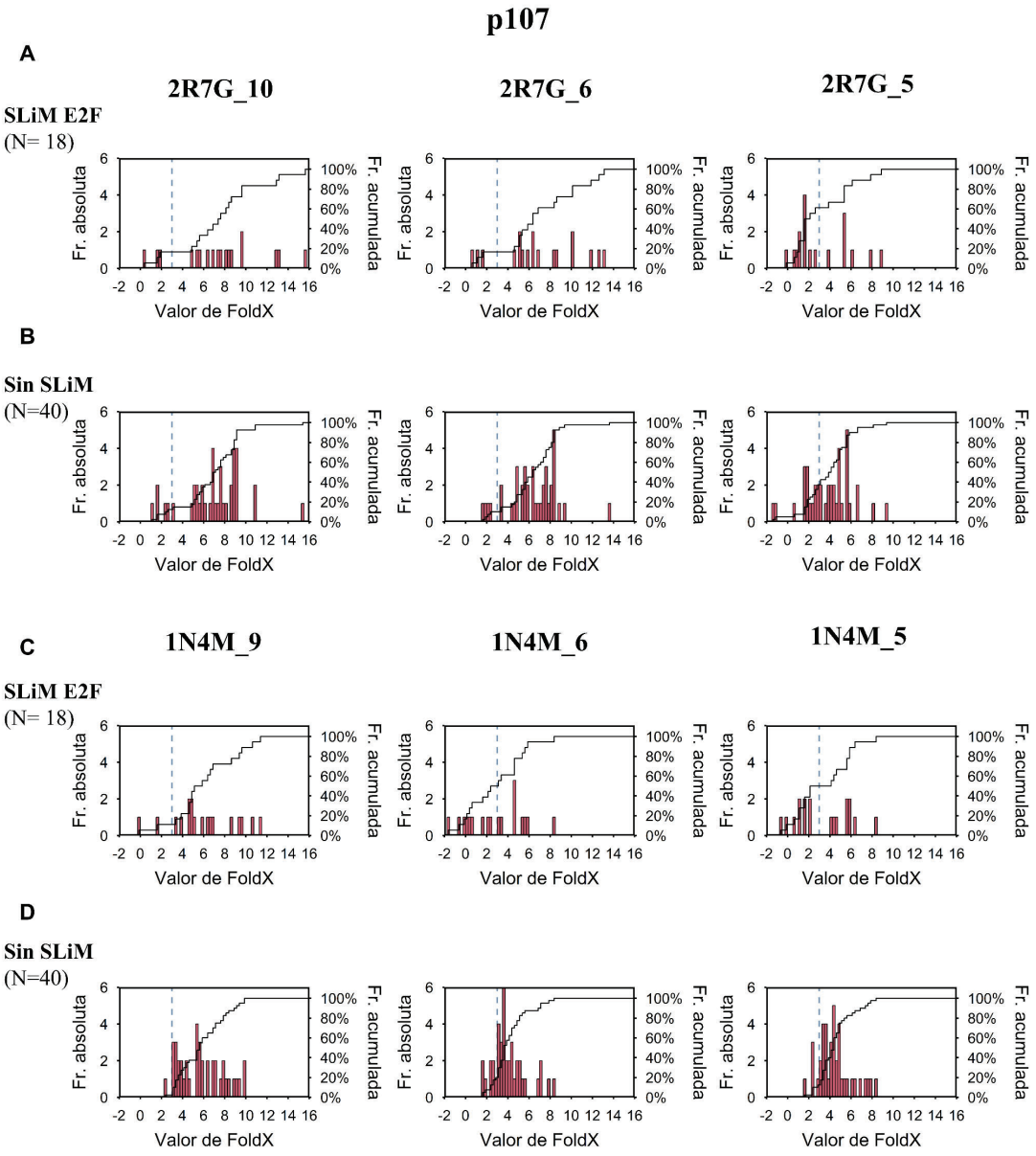


Figura S12. Péptidos *hit* de p107 escaneados con variantes de 2R7G y 1N4M. Distribución de valores FoldX de péptidos para matrices 2R7G_10, 1N4M_9 (izquierda); 2R7G_6, 1N4M_6 (medio); 2R7G_5 y 1N4M_6 (derecha). Eje Y izquierdo: Frecuencia absoluta de instancias, Eje Y derecho: porcentaje de frecuencia acumulada de instancias. La línea vertical punteada señala el valor de umbral utilizado para establecer comparaciones. **A:** Péptidos con SLiM E2F detectado escaneados con variantes de 2R7G. **B:** Péptidos sin SLiM E2F detectado escaneados con variantes de 2R7G. **C:** Péptidos con SLiM E2F detectado escaneados con variantes de 1N4M. **D:** Péptidos sin SLiM E2F detectado escaneados con variantes de 1N4M.