



UNIVERSIDAD NACIONAL DE LA PLATA
FACULTAD DE CIENCIAS EXACTAS

Implementación de una herramienta bioinformática para el diseño de secuencias linker

Trabajo final de Laboratorio de Procesos Biotecnológicos

Ignacio Eguinoa

Director: Dr. Ignacio E. Sánchez

La Plata, 2015

Resumen

La construcción exitosa de proteínas multidominio de fusión requiere una secuencia linker para unir covalentemente los dominios elegidos. Usualmente se requiere que esta secuencia no posea características funcionales que interfieran y que la misma adopte una conformación flexible, permitiendo a los dominios globulares moverse libremente. Los linkers naturales no siempre son flexibles y no pueden ser considerados inertes. Además, la diversidad de secuencias disponibles en las bases de datos de linkers no suele cubrir las propiedades requeridas por el proceso de ingeniería de proteínas. De esta forma, una aproximación racional podría ayudar en el diseño de linkers.

Este trabajo presenta un método que permite al usuario generar linkers *de novo* a partir de una secuencia random o de una secuencia impuesta por el usuario. La secuencia inicial se evalúa en busca de regiones estructuradas usando IUPRED, TMHMM, TANGO, PASTA, WALTZ, y también analizando determinantes secuenciales para la formación de fibrillas amiloides.

Se buscan posibles sitios funcionales utilizando BLAST, patrones secuenciales de las bases de datos ELM y PROSITE, y usando ANCHOR para detectar elementos de reconocimiento molecular. La carga neta y la absorción UV también pueden ser evaluadas según los requerimientos del usuario. Las características estructurales y funcionales no deseadas se mapean a cada posición de la secuencia y se calcula el total de características no deseadas. Se proponen mutaciones puntuales de forma iterativa para quitar todas las características estructurales y funcionales. Las mutaciones son aceptadas si decrece el número total de características no deseadas. Si la mutación resulta en un incremento del número de características no deseadas, la decisión se basa en una aproximación de Monte Carlo. Este método utiliza un parámetro beta para definir la probabilidad de aceptación de un determinado cambio en el número de características no deseadas, donde un valor mayor de beta se asocia a una mayor probabilidad de aceptación.

Se probaron valores de beta en el rango 0.1 a 2.5, usando secuencias iniciales random ($n=3$) y naturales ($n=3$) de largo 30. El algoritmo pudo encontrar linkers

apropiados en cada caso. El tiempo de ejecución fue más corto para valores menores de beta, con un estancamiento en el orden de los minutos por debajo de beta = 2.0. De este rango, el valor 1.0 se eligió como estándar para el método.

Diferentes ejecuciones iniciadas a partir de un conjunto compuesto por 36 secuencias random y naturales en total, con longitudes variables entre 5 y 50 residuos, también pudieron encontrar linkers apropiados en cada caso. El tiempo de ejecución se incrementó con la longitud de secuencia en una forma aproximadamente lineal.

Finalmente, se analizaron un total de 74 resultados obtenidos usando la composición de aminoácidos de UniProtKB/Swiss-Prot para las mutaciones, e iniciando a partir de una única secuencia inicial. Se encontró una alta diversidad en el conjunto de resultados. Sin embargo, este mostró una identidad remanente con la secuencia inicial. Las frecuencias de aminoácidos encontradas en el conjunto de resultados no muestran diferencias significativas con la composición aplicada durante las mutaciones. Esto permite apreciar la capacidad del método para proveer un conjunto diverso de diseños, a la vez que se minimiza el costo metabólico asociado. La composición es entonces incorporada por defecto en el método.

Usando los parámetros evaluados, el método puede encontrar linkers apropiados en un corto tiempo de ejecución. Se interpreta que el espacio de secuencias linkers apropiadas es una gran fracción del espacio completo de secuencias, mientras que el espacio de secuencias que se predicen con características estructurales y funcionales es relativamente pequeño. Como un paso hacia el desarrollo de un servidor web, el método se pone a disposición bajo el nombre de PATENA incluyendo el conjunto de parámetros estándar evaluados.

Abstract

The successful construction of multidomain fusion proteins requires a linker sequence to covalently join the selected domains. Usual requirements for this sequence are to lack any interfering functional feature and the adoption of an extended conformation, allowing globular domains to move freely. Natural linkers are not always flexible and cannot be considered inert. Furthermore, the diversity of sequences available in linker databases does not usually fill the properties required by the protein engineering process. Hence, a rational approach could aid linker design.

This work presents a method that allows the user to generate linkers de novo from a random sequence or starting from a user input sequence. The initial sequence is evaluated for structured regions using the algorithms IUPRED, TMHMM, TANGO, PASTA, WALTZ, and also scanning for sequence determinants of amyloid fibril formation.

Putative functional sites are searched using BLAST, sequence patterns in the ELM and PROSITE databases, and using ANCHOR to detect molecular recognition elements. Net charge and UV absorption can also be evaluated at the user's request. Undesired structure and functional features are mapped to each sequence position, and the total number of undesired features is calculated. Point mutations are iteratively proposed in order to remove all structural and functional features. The mutation is accepted if the total number of undesired features decreases. If the mutation results in an increased number of undesired features, the decision is based on a Monte Carlo approach. This method uses a beta parameter to define the probability of acceptance for a given change in the number of undesired features, where higher beta values are associated with higher probability of acceptance.

We tested beta values ranging from 0.1 to 2.5, using random ($n=3$) and natural ($n=3$) starting sequences of length 30. The algorithm found a suitable linker in every case. The execution time was shorter for smaller beta values, with a plateau below $\beta = 2.0$ in the minutes timescale. From this range, the value 1.0 was chosen as standard for the method.

Different executions starting from an input set comprising a total of 36 random

and natural sequences, with lengths varying from 5 to 50 residues also found a suitable linker in every case. The execution time increased with sequence length in an approximately linear manner.

Finally, we analyzed a set of 74 results obtained after using UniProtKB/Swiss-Prot amino acid composition for mutations and starting from a unique input sequence. A high degree of diversity is found in the set of results. Nevertheless, it still shows a remaining identity with the initial sequence. The amino acid frequencies found in this set of results show no significant difference with the composition applied during mutations. This allows to assess the capacity of the method to provide a diverse set of designs, while minimizing the associated metabolic cost. The composition is then incorporated to the method as default.

Using the evaluated parameters, the method can find suitable protein linkers in a short execution time. We interpret that the space of suitable linker sequences is a large fraction of the whole sequence space, while the space of sequences with predicted structural or functional features is relatively small. As a step towards the development of a web server, the method is made available under the name of PATENA including the standard set of evaluated parameters.

Agradecimientos

Prometo ser breve...

A Nacho por aceptar la dirección de esta tesis, pero más que nada por enseñarme todo para dar mis primeros pasos como investigador. Me llevo mucho más que lo que se puede leer en este trabajo.

A toda la gente de LFP por los consejos y por bancarse las miles de exposiciones que hice presentando este trabajo.

A Silvina y Gustavo, no sólo por aceptar ser jurados de este trabajo sino también por todo lo que me enseñaron de bioinformática. En cada problema que encaro me doy cuenta lo valioso que es todo lo aprendido en su materia.

A mis viejos y mis hermanos por bancarme en todo, aunque todavía no entiendan muy bien que es lo que hago y no se acuerden el nombre de la carrera.

A mis amigos del colegio, que durante la carrera se aguantaron que muchas veces no esté presente porque “estoy estudiando, tengo que rendir!” o “estoy escribiendo la tesis les juro que falta poco para recibirme!”.

A mis amigos de la facu, porque sin ellos el paso por la facu no hubiera sido lo mismo.

A todos los profesores que me dejaron algo.

A esta Universidad y en particular a esta facultad por la calidad de educación recibida.

Índice general

Resumen	2
Abstract	4
Agradecimientos	6
1. Introducción	10
1.1. Estructura de este trabajo	10
1.2. Módulos en proteínas	11
1.2.1. Definición inicial de una proteína	11
1.2.2. Estructura modular de las proteínas	12
1.2.3. Módulos proteicos	13
1.2.3.1. Proteínas globulares	13
1.2.3.2. Proteínas de membrana	18
1.2.3.3. Proteínas intrínsecamente desordenadas	20
1.2.3.3.1. Propiedades conformacionales	20
1.2.3.3.2. Propiedades secuenciales	22
1.2.3.3.3. Actividades biológicas	23
1.2.3.4. Agregados proteicos	26
1.2.3.5. Continuo resultante de estructuras, secuencias y actividades	30
1.3. Ingeniería de proteínas	31
1.3.1. Ingeniería de proteínas modulares	31
1.3.1.1. Conceptos generales	31
1.3.1.2. Ejemplos	33
1.3.2. Ingeniería de secuencias linker	35
1.3.2.1. Aspectos de diseño	35
1.3.2.1.1. Propiedades conformacionales	35
1.3.2.1.2. Elementos biológicamente funcionales	37
1.3.2.1.3. Otras propiedades	37

1.3.2.2.	Linkers naturales	38
1.3.2.2.1.	Características	38
1.3.2.2.2.	Utilización en diseños de proteínas químéricas	40
1.3.2.3.	Diseño racional	41
1.3.2.3.1.	Diseños y conceptos comunes	41
1.3.2.3.2.	Algoritmos existentes	42
1.4.	Objetivos	44
2.	Herramienta desarrollada	46
2.1.	Fundamentos del método utilizado	46
2.1.1.	Detección de propiedades a partir del análisis secuencial	46
2.1.2.	Espacio de secuencias buscadas	47
2.2.	Esquema general de la implementación	48
2.2.1.	Secuencia inicial	49
2.2.2.	Evaluación	50
2.2.3.	Mutación	51
3.	Análisis de las secuencias	54
3.1.	Propiedades conformacionales	55
3.1.1.	IUPred: Análisis de tendencia al desorden	56
3.1.2.	TMHMM: Secuencias transmembrana	58
3.1.3.	Tango	60
3.1.4.	PASTA	62
3.1.5.	Waltz	64
3.1.6.	Determinantes secuenciales de fibras amiloïdes	66
3.2.	Elementos biológicamente funcionales	67
3.2.1.	BLAST	68
3.2.2.	Prosite	70
3.2.3.	ELM	71
3.2.4.	Limbo: Interacción con chaperonas	73
3.2.5.	ANCHOR: predicción de MoREs	75
3.3.	Otros propiedades evaluadas	77
3.3.1.	Carga neta de la secuencia	78
3.3.2.	Absorción UV	79
4.	Manual de uso	80
4.1.	Instalación	80
4.2.	Parámetros de ejecución	81

4.2.1.	Secuencia inicial	81
4.2.1.1.	Secuencia inicial random	81
4.2.1.2.	Secuencia inicial definida por el usuario	81
4.2.1.3.	Secuencias flanqueantes	81
4.2.2.	Composición de la secuencia	82
4.2.2.1.	Composición estándar	82
4.2.2.2.	Composición definida por el usuario	82
4.2.3.	Definición del parámetro Beta	82
4.2.4.	Evaluación de la secuencia	83
4.2.4.1.	Evaluación de carga neta	83
4.2.4.2.	Evaluación de silente en UV	83
4.2.5.	Condición de finalización	83
4.2.6.	Formato y detalles del resultado	84
5.	Evaluaciones y análisis de resultados	85
5.1.	Parámetros de ejecución	85
5.2.	Estimación del parámetro β	85
5.3.	Análisis detallado de la ejecución	87
5.4.	Dependencia con la longitud de la secuencia	91
5.5.	Análisis de diseños resultantes	91
6.	Conclusiones y trabajo a futuro	96
	Bibliografía	97

Capítulo 1

Introducción

1.1. Estructura de este trabajo

Este primer capítulo introductorio está dividido en dos secciones que presentan los aspectos necesarios para comprender el trabajo realizado. Este se centra en el diseño de nuevas secuencias linkers con propiedades conformacionales definidas, restringiendo cualquier actividad biológica y elemento estructural no deseado.

En la primera sección (1.2) se describe el amplio panorama conformacional y funcional que pueden presentar las proteínas, lo que da una idea de cuales son las elementos proteicos que pueden encontrarse en la naturaleza. En la sección siguiente (1.3) se desarrolla el tema de ingeniería de proteínas químéricas, introduciendo el problema de cómo y por qué diseñar nuevas secuencias linker.

En el capítulo 2 se describe la implementación desarrollada para el diseño de secuencias linker, indicando los fundamentos y detalles del método.

En el capítulo 3 se detalla los distintos métodos utilizados para evaluar propiedades de interés sobre la secuencia, describiendo el objetivo de cada evaluación y los fundamentos de cada método.

En el capítulo 4 se documentan todos los aspectos necesarios para poder hacer uso de la herramienta.

En el capítulo 5 se muestran, en primer lugar, las evaluaciones realizadas para obtener los parámetros óptimos de ejecución. Además, se evalúan distintas propiedades de la herramienta resultante y su ejecución analizando, luego, los diseños resultantes.

Finalmente, el capítulo 6 contiene las discusiones relevantes, conclusiones y trabajo a futuro.

1.2. Módulos en proteínas

1.2.1. Definición inicial de una proteína

Las proteínas son polímeros de aminoácidos (fig. 1.1) y, como tales, sus propiedades conformacionales y funcionales están asociadas a la secuencia específica que las define. Es decir, las diferentes combinaciones de aminoácidos les asignan distintas capacidades de interacción intramoleculares y con el contexto en el que se encuentran, otorgándole características únicas a cada secuencia. La diversidad de aminoácidos existentes permite una gran cantidad de combinaciones, lo que resulta en un perfil de propiedades conformacionales y funcionales muy amplio.

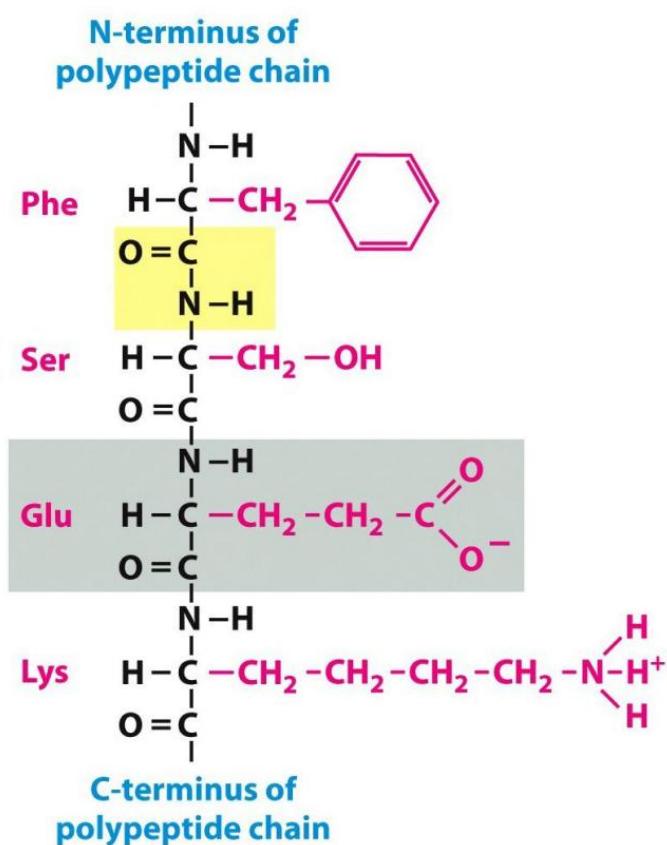


Figura 1.1: Estructura química de un polipéptido

Por definición, una proteína puede estar compuesta por cadenas lineales formadas por cualquier secuencia de aminoácidos. Sin embargo, las proteínas se originan en la naturaleza como producto de los sistemas biológicos, por lo tanto, la ocurrencia de las proteínas naturales no representa la generación de cadenas peptídicas al azar. De esta forma, la ocurrencia de secuencias en proteínas naturales está ajustada a los requerimientos del sistema y al mecanismo evolutivo subyacente. El resultado de

esto es un conjunto complejo de proteínas, donde cada una tendrá características únicas que determinan su actividad biológica dentro del sistema. Basándose en estas condiciones, las proteínas naturales quedan mejor definidas si se describen las propiedades secuenciales, estructurales y las actividades biológicas asociadas.

Cada proteína, entonces, quedará definida por el conjunto de propiedades que la describen, no sólo como polímero de aminoácidos sino también como módulo del sistema biológico (perfil de conformaciones en el contexto celular/sistémico, funcionalidades, interacciones, adaptaciones a cambios en el entorno, etc).

Si bien cada proteína tendrá algunas propiedades individuales únicas, como por ejemplo su secuencia de aminoácidos, el conjunto de proteínas naturales puede clasificarse agrupándolas de acuerdo a propiedades secuenciales y conformacionales. Cada clase de proteínas estará asociada a un cierto perfil de actividades biológicas que se derivan de sus propiedades. En las próximas secciones de este capítulo, cuando se describa alguno de los elementos proteicos naturales, se detallarán las propiedades secuenciales y conformacionales asociadas y el perfil de actividades biológicas que realizan.

1.2.2. Estructura modular de las proteínas

Las proteínas naturales son usualmente modulares, conteniendo regiones definidas asociadas cada una con una subfunción específica. Esta naturaleza modular provee muchas ventajas, principalmente la posibilidad de desarrollar funciones que requieren cooperación de distintos módulos, y un incremento en la estabilidad de la estructura global. Otras ventajas incluyen la protección de los metabolitos intermedios dentro de espacios inter-dominio, aumentando la eficiencia global y previniendo la liberación al medio de intermediarios inestables de la reacción. Desde el punto de vista evolutivo, la modularidad en las proteínas incrementa la capacidad evolutiva reduciendo las restricciones necesarias para la adaptación y permitiendo que módulos preexistentes puedan funcionar en nuevos contextos para nuevos usos.

El módulo o unidad de proteína más común es el dominio [111], el cual puede adquirir distintas definiciones según las propiedades que se analicen: Podemos definir a un dominio como una unidad evolutiva independiente que puede dar una proteína mono-dominio o formar parte de una proteína multi-dominio. Cada dominio puede tener una función independiente o contribuir a la función global de una proteína multi-dominio cooperando con el resto de las unidades. Esta definición de dominio como unidad evolutiva es la que se usa como mecanismo de clasificación en la base de datos SCOP (Structural Classification of Proteins) [79].

A diferencia de esto, en CATH [84], los dominios se definen basándose en con-

ceptos puramente estructurales. De esta forma, los dominios pueden definirse como porciones de la secuencia de un polipéptido que pueden asumir una estructura tridimensional estable de forma independiente.

Los dominios, si bien son los componentes más estudiados de las proteínas, no son los únicos, existiendo otras unidades que pueden estar representadas por regiones más cortas e incluso limitarse a simples motivos secuenciales compuestos de unos pocos residuos contiguos [108]. Cada tipo de módulo tendrá sus propiedades distinguibles.

La arquitectura de una proteína, entonces, estará dada por la combinación de distintos módulos identificables por sus propiedades secuenciales, estructurales o actividades biológicas características, posiblemente unidos por regiones secuenciales que pueden no tener una función individual pero que en conjunto hacen a la actividad biológica global de la proteína.

En la figura 1.2 se muestra esta composición modular en un conjunto de proteínas efectoras de bacterias. La representación gráfica de dominios y motivos con colores permite identificar módulos individuales que se repiten en distintas proteínas.

1.2.3. Módulos proteicos

1.2.3.1. Proteínas globulares

Las proteínas globulares son los compuestos proteicos más estudiados. En términos generales, estas proteínas se caracterizan por adoptar en condiciones nativas una conformación plegada compacta, dándole un aspecto globular y haciéndola soluble en el medio acuoso de la célula.

El hecho que sean, probablemente, las proteínas más estudiadas tiene está asociado a estas propiedades estructurales básicas, al ser solubles y con una estructura determinada y relativamente estable, fueron las primeras que pudieron ser cristalizadas para luego resolver su estructura mediante difracción de rayos X. La anotación de las estructuras que se resolvían permitió conocer más sobre las propiedades conformacionales de estas, mostrando que las proteínas globulares poseen una gran cantidad de estructuras secundarias interaccionando entre sí, unidas por regiones de la secuencia con estructura flexible no regular (ver figura 1.3).

Las distintas estructuras secundarias se unen entre sí mediante interacciones hidrofóbicas que involucran las cadenas laterales apolares, y por puentes de hidrógeno y a través de interacciones iónicas entre aminoácidos cargados, dando lugar a la estructura compacta del plegamiento globular. Por lo tanto, el plegamiento adoptado por las estructuras secundarias (la llamada estructura terciaria) está intimamen-

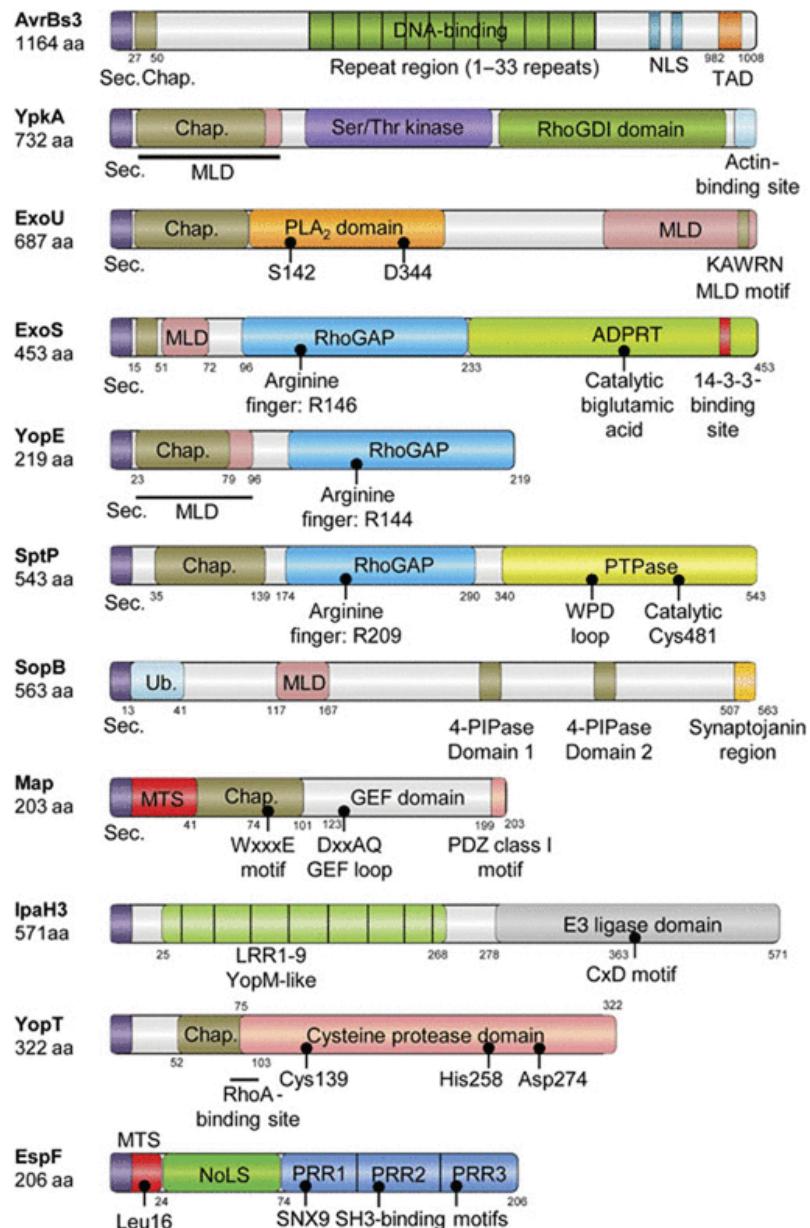


Figura 1.2: **Arquitectura de un conjunto de proteínas efectoras de bacterias.** Algunos dominios identificables son: Sec. = secretion domain; Chap = chaperone-binding domain; PRR = proline-rich repeats. Se identifican otros tipos de módulos, por ejemplo motivos que pueden estar en cualquier parte de la secuencia. Figura extraída de [36]

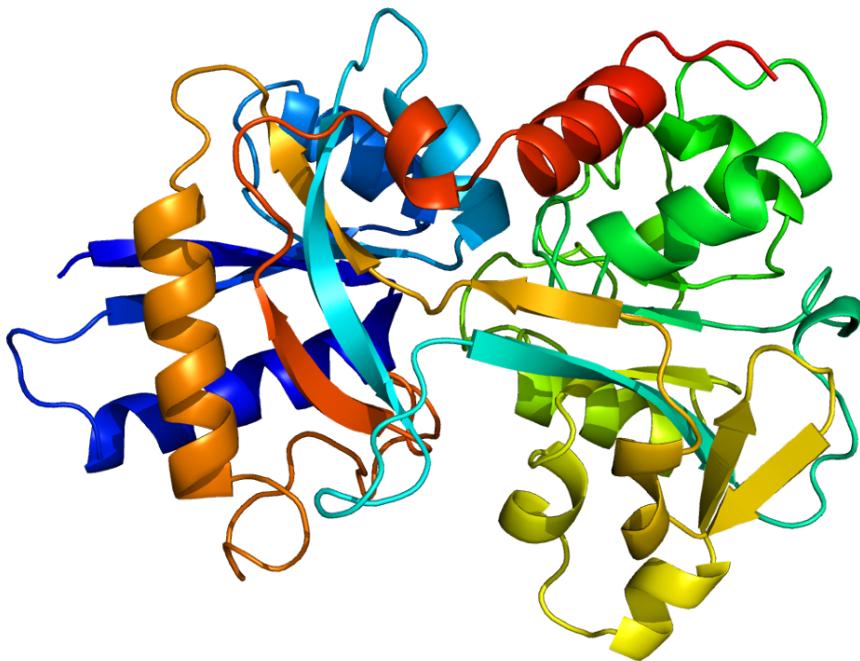


Figura 1.3: **Proteína con estructura globular.** Imagen de: PDB 1A8E (Transferrin)

te asociado a las propiedades fisicoquímicas de la secuencia. En este plegamiento compacto, los aminoácidos más hidrofóbicos quedan alojados principalmente en el interior de la estructura, formando un núcleo hidrofóbico, y la superficie globular tiene mayoría de residuos polares los cuales tienen la capacidad de interaccionar favorablemente con el solvente, brindándole solubilidad en el entorno celular.

La estructura globular puede estar conformada por más de una cadena polipeptídica, en cuyo caso las diferentes subunidades se unen mediante las mismas interacciones que generan el plegamiento en una cadena. De esta forma, el plegamiento individual de una cadena puede contener aminoácidos con cadena lateral apolar en alguna región de su superficie, pero esta región no queda expuesta al entorno sino que se complementa con la superficie hidrofóbica de otra proteína. Estas interacciones entre cadenas distintas mantienen la estructura global del complejo.

Todas estas propiedades secuenciales diferenciales a lo largo de la estructura se asocian con el paradigma clásico de la biología estructural, el cual afirma que la secuencia primaria de una proteína contiene la codificación para que esta se pliegue adoptando una estructura tridimensional determinada y que, además, la conformación final se corresponde con un mínimo global sobre el perfil de energía libre asociado al polipéptido.

Teniendo en cuenta que adoptar una estructura determinada en un polipéptido

implica un gran costo entrópico, y que las interacciones que describimos como estabilizadoras de este plegamiento son principalmente débiles, es de esperar que una característica de las proteínas globulares sea la capacidad intrínseca de su secuencia para formar gran cantidad de interacciones que permitan estabilizar la estructura plegada.

El paradigma estructura-función original afirmaba que la función específica de una proteína está determinada por su estructura tridimensional, la cual es única y rígida. El origen de este paradigma es el modelo de llave-cerradura propuesto por Emil Fisher en 1894, que nace de la necesidad de explicar la especificidad encontrada en el funcionamiento de diversas enzimas similares por sus correspondientes sustratos. La hipótesis utilizada para explicar esto era que la estructura de la proteína (en este caso la enzima) era perfectamente complementaria a la del sustrato correspondiente, en un sistema similar al de una llave en una cerradura. Por un largo período de tiempo, entonces, el modelo de llave-cerradura y el paradigma asociado de estructura-función se mantuvieron como teorías aceptadas, compatibles con muchas de las estructuras que se resolvieron mediante difracción con rayos X.

Contrario a este modelo, en 1958 Koshland sugiere el modelo de encaje inducido, basado en observaciones de enzimas que podían aceptar ligandos con diferentes estructuras. Por lo tanto, la proteína debía poseer cierta flexibilidad que le permitiera esta función. Posteriores evidencias mostraron que esta flexibilidad era parte del mecanismo funcional y que podía implicar cambios conformacionales que van desde variaciones en rotámeros de las cadenas laterales, hasta reordenamientos conformacionales muy relevantes como en el caso de proteínas alostéricas. Estas variaciones aún podían ser interpretadas dentro del modelo establecido por el paradigma estructura-función, el cual fue levemente adaptado. El requerimiento necesario para que la proteína cumpla su función era, entonces, una estructura definida, aunque dinámica. Por un tiempo todo pareció indicar que una estructura definida, aunque dinámica, era la base del mecanismo funcional.

Si bien esta idea es una simplificación, dado que aún las proteínas más estables son sistemas dinámicos con distintos grados de flexibilidad, este modelo se adapta a las funciones asociadas a proteínas globulares [61]. En éstas, el ensamblaje conformado por los estados accesibles en condiciones fisiológicas está limitado a un conjunto de estructuras que sólo difieren en leves desviaciones del promedio del ensamblaje. La estructura resultante que se observa de la cristalización es, en realidad, un promedio del ensamblaje de conformaciones.

La dinámica de estas estructuras está dada por movimientos que van desde fluctuaciones rápidas de pequeña amplitud (del orden de Å) hasta cambios relativamente

lentos ($\mu s - s$) que involucran modificaciones estructurales relevantes (plegamiento y movimientos asociados a la función). Todos los movimientos son resultantes de rupturas y formaciones de interacciones provocadas por fuerzas de interacción las cuales, debido a su naturaleza débil, pueden romperse por efecto de la energía térmica, aún a temperatura ambiente. En general, todos estos movimientos están asociados al mecanismo por el cual ejercen su función las proteínas globulares.

Sabemos ahora que la funcionalidad emergente de los dominios globulares depende no sólo de la estructura promedio del ensamble sino también de los aspectos dinámicos de ésta. Todas estas características imponen requerimientos al mecanismo evolutivo ya que, para mantener la funcionalidad intrínseca de la proteína, se deben mantener los determinantes secuenciales y estructurales asociados al mecanismo funcional. Esto impone ciertos requerimientos sobre la secuencia y/o estructura para la evolución de la proteína, que resulta en una similitud entre proteínas homólogas.

Estos conceptos han sido de gran ayuda en el estudio y clasificación de nuevas proteínas. Conociendo la estructura asociada a una proteína se puede inferir la función que desarrolla a partir de la búsqueda de proteínas con estructuras similares cuya función sea conocida. Este mismo método se puede aplicar sobre dominios individuales ya que, como vimos, actúan como unidades estructurales y evolutivas individuales, mostrando funcionalidades independientes.

Incluso cuando no se conoce la estructura, los requerimientos conformacionales imponen indirectamente restricciones sobre la secuencia, ya que la estructura es resultado directo de esta. De tal manera, la comparación de nuevas secuencias con las de otras cuya función es conocida permite inferir propiedades funcionales a partir de la similitud secuencial, lo que facilita la anotación de elementos funcionales, aún cuando no se conoce la estructura de la proteína.

Estos conceptos asociados de propiedades secuenciales, estructurales y funcionales, unidos por el mecanismo evolutivo son la base de gran parte del conocimiento actual sobre proteínas globulares.

El perfil de actividades biológicas encontradas en las proteínas globulares naturales está directamente relacionado con las propiedades conformacionales de estas. En general, las funciones más encontradas están asociadas con actividades enzimáticas, donde las proteínas se encuentran generalmente solubles en el medio acuoso de la célula. La estructura tridimensional definida y gran tamaño que pueden alcanzar las proteínas globulares también les permite intervenir en actividades que requieren interacciones específicas y estables con otras moléculas. Ejemplo de esto son los anticuerpos y las proteínas que intervienen en la replicación y reparación del ADN. El gran tamaño y la solubilidad de las proteínas globulares les permiten actuar también

como transportadores de otras moléculas con baja solubilidad en medio acuoso.

La flexibilidad estructural permite que se experimenten cambios conformacionales considerables en la estructura terciaria. Dado que las propiedades conformacionales están intimamente ligadas a la funcionalidad, estos cambios, que pueden ocurrir como producto de la unión de ligandos o modificaciones sobre químicas dinámicas sin alterar la secuencia, permiten ejercer efectos regulatorios sobre la actividad de la proteína.

1.2.3.2. Proteínas de membrana

Las proteínas de membrana se definen como cualquier proteína que interacciona con la membrana lipídica de la célula. Algunas de estas proteínas están unidas sólo a la superficie de la membrana, mientras que otras tienen regiones que se insertan en el núcleo hidrofóbico. Se puede hacer, entonces, una primera clasificación en 2 grupos: proteínas integrales y periféricas.

Las proteínas extrínsecas (o periféricas) no interaccionan con el núcleo hidrofóbico y tienen propiedades estructurales y secuenciales similares a las proteínas globulares, pero se mantienen en constante interacción con la membrana mediante distintos mecanismos. Algunas interaccionan indirectamente mediante la unión a proteínas integrales de membrana, otras lo hacen a través de la interacción con los grupos polares que componen las cabezas de los fosfolípidos en la membrana. Otro conjunto de proteína periféricas atraviesan procesos de modificaciones post-traduccionales que las unen covalentemente a cadenas carbonadas, las cuales pueden insertarse en el núcleo hidrofóbico y así mantener a la proteína en constante interacción con la membrana. En general, esta clase de proteínas tiene actividades biológicas asociadas a la unión a la membrana, por ejemplo transducción de señales extracelulares, están involucradas en la cadena de transporte de electrones, etc.

El otro conjunto de proteínas de membrana son las proteínas integrales o intrínsecas [112], que tienen uno o más segmentos insertos en la bicapa lipídica y, en la mayoría, la región insertada atraviesa la membrana (proteínas transmembrana), teniendo dominios intra y extra celulares como parte de la arquitectura global. A partir de estos conceptos generales se puede deducir que esta clase de proteínas tendrá una arquitectura/topología particular, donde los diferentes dominios tendrán propiedades secuenciales y estructurales distinguibles dependiendo si se encuentran en el interior de la membrana o si son intra(o extra) celulares.

Los segmentos que atraviesan la membrana permiten hacer una nueva clasificación, así, las proteínas transmembrana se pueden agrupar en dos conjuntos según la estructura que les permite atravesar el núcleo hidrofóbico (ver figura 1.4). Un primer

conjunto de proteínas, las porinas, contienen una estructura característica compuesta por hebras- β en una conformación característica con forma de barril que atraviesa la membrana, formando una expansión similar a un poro [115]. Los aminoácidos de la secuencia son generalmente polares pero, a diferencia de las proteínas globulares típicas, los grupos laterales que están en contacto con la cara externa del barril son hidrofóbicos, y son los que están en contacto con los grupos lipídicos, mientras que las cadenas laterales hidrofílicas se encuentran mirando hacia el interior del poro. La actividad biológica de estas proteínas se deriva directamente de esta estructura, las porinas forman canales en la membrana a través de los cuales distintos metabolitos solubles en agua pueden atravesarla.

El otro conjunto está dado por aquellas proteínas que contienen hélices- α atravesando una o más veces la membrana. Esta estructura secundaria es la más común entre los segmentos transmembrana ya que permite satisfacer internamente todos los puentes de hidrógeno, sin dejar grupos polares expuestos hacia fuera de la hélice, y por lo tanto, en contacto con el núcleo hidrofóbico de la membrana

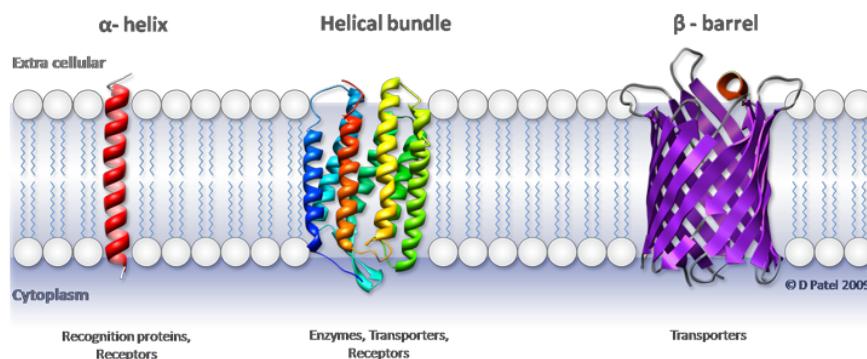


Figura 1.4: **Tipos de dominios intramembrana:** dominios con una sola hélice- α atravesando la membrana (izquierda), dominios múltiples hélices- α atravesando la membrana e interaccionando entre si (centro), dominios compuestos por láminas β formando una estructura similar a un barril

Los dominios transmembrana tienen propiedades secuenciales particulares caracterizadas por aminoácidos con cadenas laterales hidrofóbicas que les permiten interaccionar con el núcleo de la membrana [80]. Un ejemplo común de este tipo de proteínas es la glicoforina, que contiene un segmento con estructura de hélice- α compuesto totalmente de residuos hidrofóbicos. Esta hélice hidrofóbica está, además, rodeada por segmentos flanqueantes conteniendo aminoácidos cargados que interactúan con las cabezas polares de los fosfolípidos y previenen que la proteína se deslice, manteniéndola inserta en la membrana. Dentro de este último conjunto existe una gran cantidad de proteínas que contienen múltiples hélices- α transmembrana. Estás proteínas generalmente están conformadas por 2 hélices formando una

estructura tipo coiled-coil, o por 7 hélices, en un arreglo característico de proteínas como la rodopsina [85].

Estas particularidades en la secuencia permiten distinguirlas fácilmente de, por ejemplo, las proteínas globulares. Además, a partir de estos módulos transmembrana y extra/intra celulares se pueden clasificar estas proteínas según la arquitectura o topología global, de acuerdo al número y tipo de segmentos transmembrana, etc.

1.2.3.3. Proteínas intrínsecamente desordenadas

1.2.3.3.1. Propiedades conformacionales

El conjunto de proteínas desordenadas (IDRs/IDPs = intrinsically disordered regions/proteins, a partir de ahora) se caracteriza por la tendencia a adoptar conformaciones extendidas con bajo contenido de estructura secundarias [54]. La principal característica distintiva de las IDPs es, entonces, su inhabilidad para plegarse en una estructura tridimensional única, aunque en ciertos casos pueden formar estructuras desordenadas pero compactas, y algunos segmentos de la secuencia pueden adoptar de manera transiente estructuras secundarias individuales (ver figura 1.5). Esta leve estructura residual (latente o transiente) es muy importante y representa gran parte de las capacidades funcionales en IDRs/IDPs.

A pesar de esta tendencia desordenada intrínseca, debido a la naturaleza heteropolimérica de las proteínas, es probable que las IDPs **no** adopten conformaciones totalmente aleatorias, aún en entornos altamente desnaturalizantes. Por lo tanto, se puede ver que las IDPs poseen propiedades conformacionales complejas, y que esta excepcional heterogeneidad estructural representa también un problema crítico a la hora de estudiarlas experimentalmente y caracterizarlas [43].

Una característica resultante de estas propiedades es que las IDPs poseen conformaciones altamente dinámicas con estructuras que se interconvierten en diferentes escalas de tiempo. Estas estructuras varían desde estados completamente desordenados (polipéptidos nativos sin plegamiento) hasta estados parcialmente plegados sin estructuras secundarias definidas, o incluso ensambles desordenados pero con ciertos segmentos adoptando estructuras secundarias claramente determinadas. La distribución de estas estructuras está cambiando continuamente en el tiempo, es decir, la estructura tridimensional que podemos ver en un instante dado será diferente de la que veremos en otro momento.

Las transiciones rápidas y frecuentes entre las distintas conformaciones resultan en las observaciones experimentales en solución: un conjunto heterogéneo de conformaciones fácilmente maleable por cambios en el medio. En base a esto, se ve que una descripción detallada de IDPs requeriría conocer el ensamble de estados

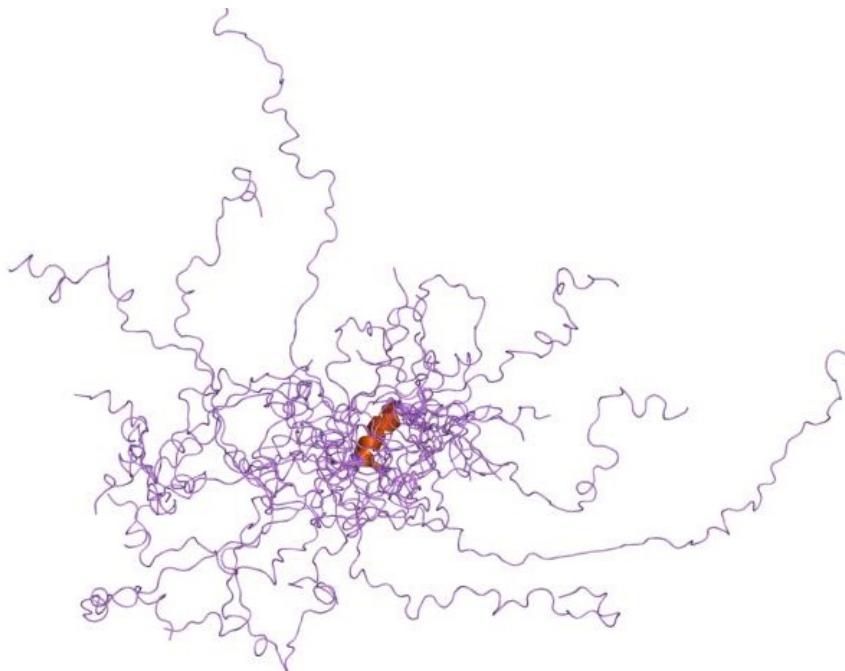


Figura 1.5: **Ejemplo de IDP.** Ensamble estructural de la proteína TSP9 (Thylakoid soluble phosphoprotein), una proteína intrínsecamente desordenada. Estructura resuelta mediante NMR. PDB: 2FFT

posibles junto con las velocidades de interconversión entre estos. Todas estas propiedades y las velocidades de las transiciones en el ensamble son difíciles de capturar experimentalmente, aún usando técnicas experimentales que no se basan en promedios de las conformaciones. Los estudios sobre este tipo de proteínas, entonces, se han enfocado en analizar los estados termodinámicamente accesibles como entidades discretas. Esta idea, si bien sabemos que es una descripción incompleta, permite modelar de forma discreta los elementos funcionales relevantes, sirviendo como base para estudios posteriores. Este modelo puede extenderse si se representan también las probabilidades de ocurrencia para distintas conformaciones individuales en el ensamble, obteniendo así un modelo más dinámico del sistema.

Los ensambles conformacionales pueden verse modificados por distintas situaciones como cambios en las condiciones del entorno, modificaciones post-traduccionales, presencia de ligandos, etc., lo que modifica no sólo el conjunto de estados accesibles, sino también la distribución de poblaciones que se encuentran en cada uno. Una propiedad particular dependiente del contexto, encontrada en los ensambles conformacionales asociados a IDPs, es la capacidad de adquirir nuevos elementos estructurales ordenados estables luego de la unión a ciertos ligandos (generalmente otras proteínas, DNA, membranas, etc), en un proceso conocido como *folding & binding* [42]. Esta propiedad se hace más relevante de investigar cuando se sabe

que, generalmente, está asociada a su funcionalidad biológica y ocurre previamente o durante la realización de esta función.

En general, sólo un segmento desordenado con características anfipáticas es el que está involucrado en el proceso de unión y plegamiento. En este caso el segmento se denomina elemento de reconocimiento (RE=recognition element). Este tipo de transiciones ante la unión a ciertos ligandos, involucrando regiones cortas de la secuencia, provee una combinación de alta especificidad y baja afinidad que hace a este tipo de elementos extremadamente útiles en procesos de señalización y regulación. Una pregunta que aparece naturalmente al intentar estudiar este proceso es si el elemento estructural ordenado es generado durante la unión al ligando o si pertenece al conjunto de elementos estructurales encontrados transitivamente en las IDR/IDPs y, al unirse, se hace parte de la conformación estable. Estas dos opciones se reflejan en los modelos de selección de conformación y el de unión y plegamiento simultáneo. Obviamente, en la realidad puede ocurrir cualquiera de estos dos mecanismos o una combinación de los dos. Esto se ve representado en la figura 1.6

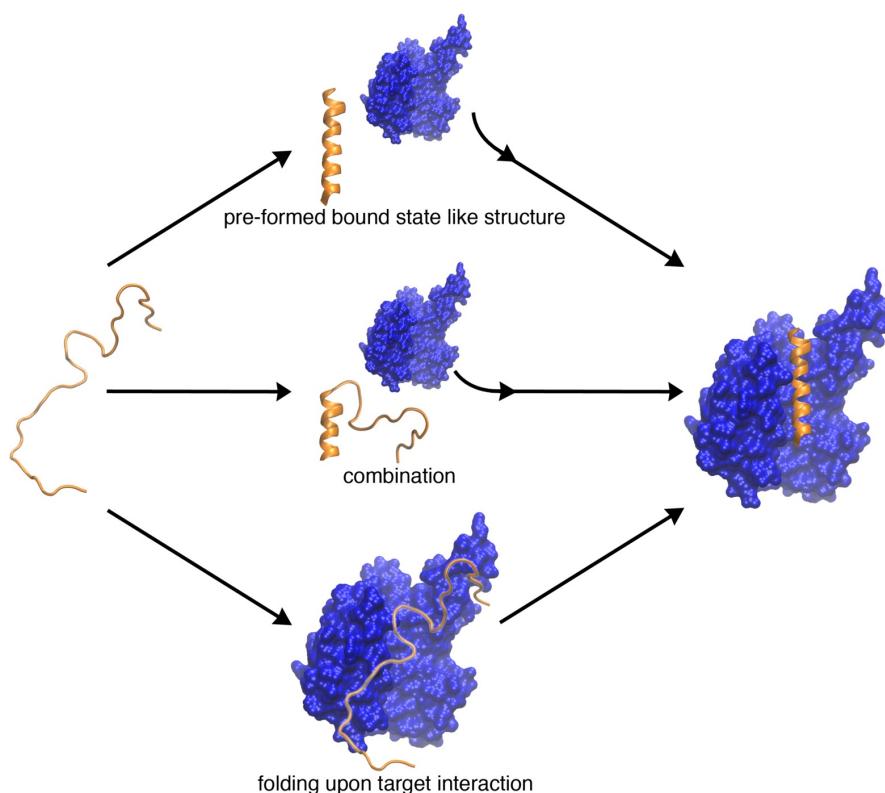


Figura 1.6: Distintos modelos de *binding & folding*

1.2.3.3.2. Propiedades secuenciales

Muchos estudios se hicieron para intentar identificar cómo las características

estructurales tan particulares de las IDPs están codificadas en la secuencia de la proteína [104]. En el caso de las proteínas globulares, la secuencia primaria codifica el proceso de plegamiento y, a través de este, la estructura única final. De forma similar, al identificar las IDPs como entidades estructuralmente diferenciables, se podría pensar que la estructura primaria de estas determina, también, su falta de estructura tridimensional definida. Esto tiene como corolario la idea que las secuencias de ambos conjuntos (proteínas ordenadas y proteínas desordenadas) deberían ser diferenciables entre sí.

La ausencia de una estructura definida en IDPs es atribuida a las características particulares de sus secuencias de aminoácidos las cuales se describen como secuencias de baja complejidad [114], enriquecidas en aminoácidos cargados y polares, junto con Glicina y Prolina, y un bajo contenido de residuos hidrofóbicos. La secuencia suele tener una alta carga neta resultante, o tener segmentos separados de residuos cada uno cargas positivas o negativas [73]. La relación entre esta combinación de baja hidrofobicidad y alta carga neta y la falta de una estructura definida puede explicarse desde el punto de vista físico: los altos valores de carga neta producen efectos de repulsión entre las cargas de los residuos, mientras que la baja hidrofobicidad previene la compactación de la estructura extendida (las interacciones hidrofóbicas son una de las principales fuerzas que intervienen en el proceso de plegamiento). De esta forma, las IDPs no poseen las propiedades secuenciales que le permitan formar suficientes interacciones para estabilizar una estructura compacta con un núcleo hidrofóbico, característica de las proteínas globulares. El resultado es una estructura con tendencia a mantener una conformación desplegada, característica de las IDPs.

1.2.3.3.3. Actividades biológicas

Una pregunta importante que surge de las propiedades estructurales es como estas desarrollan las funcionalidades asociadas y cuáles son las funciones biológicas que cumplen las IDPs. Un paso importante para entender esto es la clasificación de las IDPs conocidas de acuerdo a sus funcionalidades [106]. La figura 1.7 detalla la clasificación de funcionalidades y mecanismos asociados a IDPs junto con las proporciones de instancias encontradas para cada uno. De esta clasificación se puede ver que las funcionalidades en las IDPs emergen a partir del propio desorden intrínseco o mediante mecanismos de reconocimiento molecular.

En el primer caso, el mecanismo subyacente de funcionamiento no implica ninguna interacción sino que depende directamente de la flexibilidad y plasticidad de la cadena carbonada (*entropic chains*). En este caso, las proteínas satisfacen un conjunto de requerimientos celulares únicos en los cuales la función a cumplir es una

consecuencia directa de las propiedades conformacionales que poseen. Esta función, generalmente está asociada a secuencias linkers, cuyo objetivo es unir distintos componentes de las proteínas.

En el segundo caso, el amplio perfil de interacciones posibles permite que las IDPs participen en funcionalidades muy diversas, entre las cuales están la función de chaperonas, la modulación de actividades de otras moléculas y la neutralización de pequeños ligandos uniéndose a ellos. El mecanismo de reconocimiento e interacción es básicamente diferente al que adoptan los dominios globulares. Mientras que estos últimos han evolucionado para dar una gran variedad de plegamientos diferentes que proveen reconocimientos específicos, en IDPs el mecanismo de unión es principalmente mediado por segmentos cortos de reconocimiento [50, 81], caracterizados por su naturaleza flexible [42, 53].

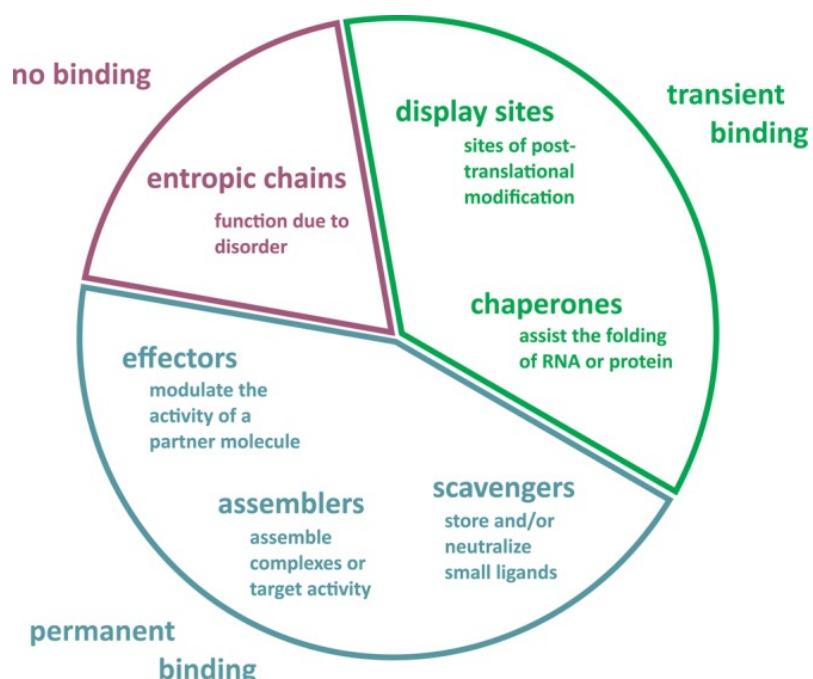


Figura 1.7: Clasificación roles y mecanismos funcionales de IDPs. Figura extraída de [106]

Dentro de estos segmentos podemos identificar a los elementos de reconocimiento [78, 83, 105], los cuales involucran regiones cortas (10-70 aminoácidos de largo) con propiedades generalmente anfipáticas, y que se caracterizan porque pueden experimentar un proceso plegado y unión (descrito previamente) que permite la interacción con ligandos.

Las IDPs pueden interaccionar también a partir de elementos en la secuencia conocidos como motivos lineales o por sus siglas en inglés, LMs, ELMs, SLiMs o MiniMotifs [108]. Estos se caracterizan por presentar interfaces de interacción muy

compactas, donde los residuos determinantes para la afinidad y especificidad están contenidos generalmente en sólo 3-11 posiciones contiguas. Como resultado del número limitado de contactos para interacciones con ligandos, los SLiMs se unen con una afinidad relativa muy baja lo que resulta en una ventaja considerable para participar en interacciones transientes, condicionales y ajustables, que son propias de los procesos regulatorios. En términos de funcionalidad, se pueden clasificar los SLiMs en dos grandes familias: los que actúan como sitios objetivo de modificaciones post-traduccionales, y los que funcionan como ligandos en la formación de complejos de interacción.

El hecho de que estén definidos sólo por un pequeño segmento secuencial tiene consecuencias relevantes en distintos aspectos biológicos de los SLiMs [31]. En particular, claramente resulta muy probable que estos aparezcan o desaparezcan mediante mutaciones. Dado que solo se necesitan un número muy limitado de mutaciones para que surgan, estos son propensos a la evolución convergente, lo que facilita la proliferación en distintos lugares del proteoma. Esto resulta en una fuente de evolución para la red de interacciones ya que se generan interfaces *de novo* en distintas proteínas. Estas mismas características dinámicas de la secuencia los hace más difíciles de estudiar que elementos funcionales más restringidos evolutivamente como los dominios globulares. Por lo tanto, a pesar de la disponibilidad de miles de secuencias de proteínas, el descubrimiento de motivos contenidos en estas es aún una tarea difícil.

A pesar que las interacciones mediante SLiMs suelen ser débiles, transientes y, posiblemente, tener baja especificidad, pueden originarse interfaces más específicas y de mayor afinidad si se obtiene cooperación de las regiones flanqueantes, si se combinan varios motivos cortos, o si se utilizan dominios desordenados con mayor longitud. Es relevante decir que las regiones de unión en IDPs suelen corresponderse más con dominios que con motivos cortos, teniendo longitudes que exceden los 20-30 residuos [25, 26, 101]. Además de esta longitud característica, tienen distintas propiedades que los distinguen como dominios: Son estructuralmente y funcionalmente independientes dentro de la proteína, pueden ser más fácilmente reconocibles mediante similitud secuencial debido a que sus secuencias están conservadas, y poseen al menos una función específica que los identifica. Al poder distinguirlos como dominios intrínsecamente desordenados (IDDs) podemos clasificarlos, entonces, como una nueva clase de elementos funcionales independientes que pueden encontrarse dentro de IDPs. Estos elementos tienen funciones muy diversas pero, generalmente, están involucrados en procesos de interacción con DNA, RNA u otras proteínas.

1.2.3.4. Agregados proteicos

En términos generales, los agregados proteicos se forman por la asociación entre proteínas, dando complejos con distintas propiedades y que pueden acumular una gran cantidad de cadenas polipeptídicas independientes. Si bien algunos agregados pequeños pueden mantenerse en solución, en general, bajo condiciones fisiológicas normales, terminan formando precipitados.

Es importante diferenciar, entre los distintos agregados insolubles, los precipitados en los que las proteínas mantienen su estructura nativa de otros agregados formados a partir de nuevas estructuras no-nativas (ver gráfico 1.8). La precipitación de conformaciones nativas es generada por la inducción de un entorno de solubilidad reducida, por ejemplo durante la precipitación isoeléctrica o mediante sulfato de amonio, que son los procesos típicos para obtener los cristales utilizados en la resolución de estructuras por difracción de rayos X. En estos casos, la reducción de la fuerza iónica o el cambio en el pH de la solución restituye los precipitados a la forma soluble. El segundo tipo de precipitados macromoleculares representa la formación de interacciones entre proteínas, principalmente formando hojas- β , que llevan a la formación de estados muy estables, reflejados en un profundo mínimo sobre el perfil de energía correspondiente. Esto hace que solo sea posible restituir la forma soluble de los componentes mediante procedimientos extremos aunque, igualmente, es difícil que las proteínas vuelvan a adoptar sus conformaciones nativas.

Este último tipo de agregados, donde se pierden las conformaciones nativas de las proteínas individuales, adquiere nuevas características estructurales distintivas. A partir de las características morfológicas de estos, se pueden distinguir esencialmente dos tipos [92]: agregados amorfos y fibrillas amiloïdes. Los primeros muestran una apariencia granular y consisten principalmente de proteínas en conformaciones desplegadas, aunque presentan ciertas regiones cierto enriquecimiento de estructuras de hoja- β plegada, las cuales mantienen las proteínas unidas formando el agregado. Por otro lado, las fibrillas amiloïdes poseen estructuras altamente ordenadas y repetitivas donde todos los polipéptidos que la componen adoptan un plegamiento en común. Estas diferencias que se ven en la conformación macromolecular son el reflejo de diferencias en las interacciones y arquitectura a nivel atómico [72].

Las fibrillas amiloïdes son, probablemente, el tipo de agregado más interesante y quizás también el más estudiado [98] ya que tiene propiedades que las hacen muy particulares, además de estar fuertemente asociados enfermedades neurodegenerativas [91]. Los depósitos encontrados en el contexto de estas enfermedades estaban conformados generalmente de un solo tipo de proteínas, aunque se encontraban *in-vivo* asociados a distintas moléculas. Los estudios más avanzados de su

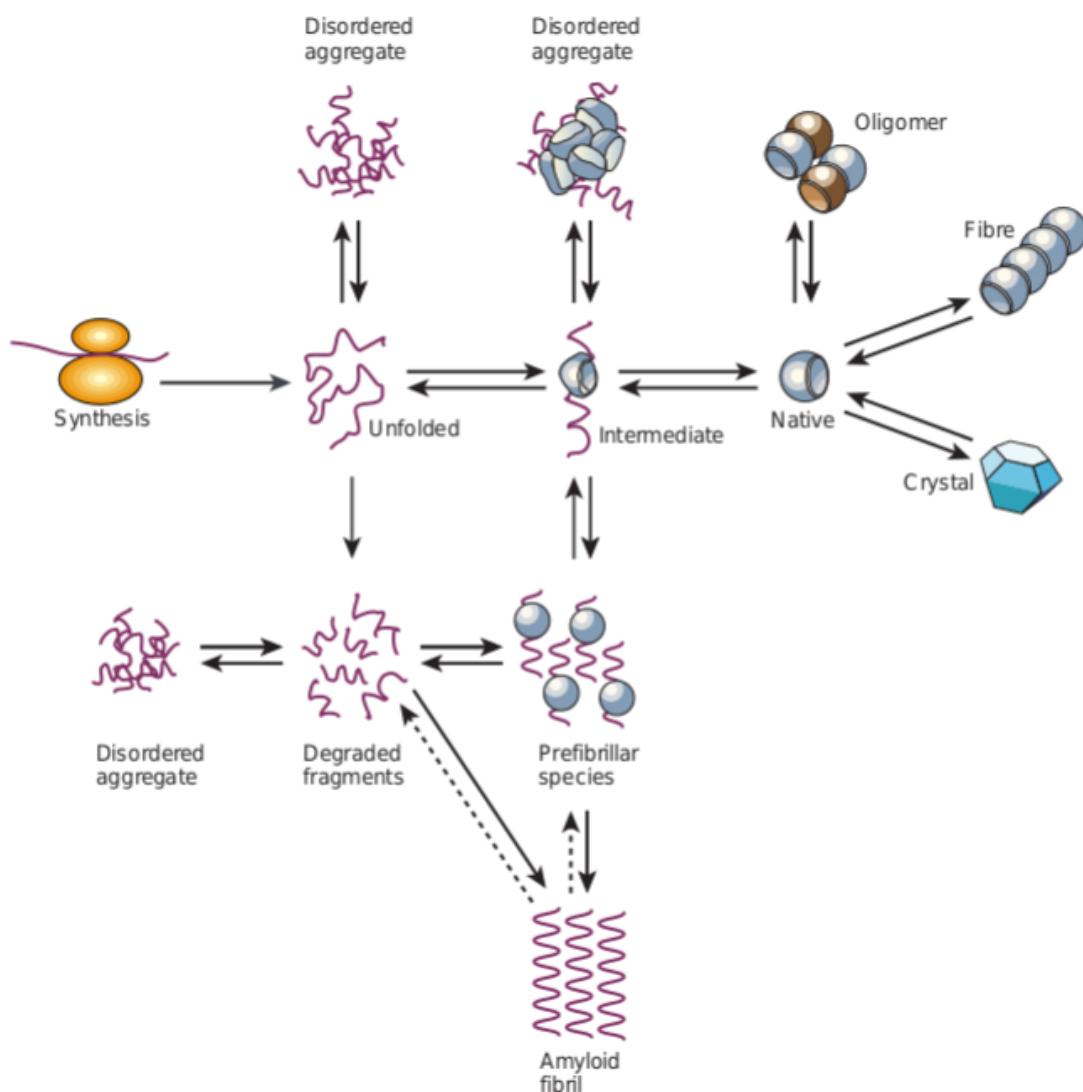


Figura 1.8: **Perfil de los distintos agregados proteícos:** Integrados dentro del perfil conformacional de las proteínas, se muestran los distintos tipos de estructuras de agregación, los intermediarios a partir de los cuales se originan y las unidades que los componen.

estructura revelaron las características altamente ordenadas, compactas, estables y no-ramificadas que poseen estos agregados y que se ven reflejados en la morfología fibrilar que puede apreciarse al verlos mediante el microscopio electrónico de transmisión. El diámetro de estas estructuras fibrilares se encuentra en el rango de los 10 nanómetros y su longitud puede alcanzar varios micrómetros. Estas fibrillas maduras pueden, además, asociarse lateralmente entre si para formar fibras.

Todas las fibrillas amiloïdes comparten una arquitectura en común compuesta de una estructura suprasecundaria de unión entre hojas- β (ver figura 1.9). Los detalles de esta arquitectura, obtenidos mediante difracción con rayos X, revelan

que las hojas- β se extienden con las caras de sus hebras enfrentadas entre si pero perpendiculares al eje de la fibrilla que forman [82]. Esta arquitectura de hojas- β cruzadas permite, como se mencionó antes, la formación de un continuo de puentes de hidrógeno que provee a estas componentes de gran estabilidad.

Además de la estructura ordenada y compacta, se caracterizan por poseer una alta estabilidad cinética, probablemente debido a su estructura de puentes de hidrógeno altamente organizada. Por lo tanto, una vez formados estos agregados pueden mantenerse durante largos períodos de tiempo, funcionando como núcleo para la agregación de más cantidades de la misma proteína y permitiendo la formación de grandes depósitos en la célula/tejido.

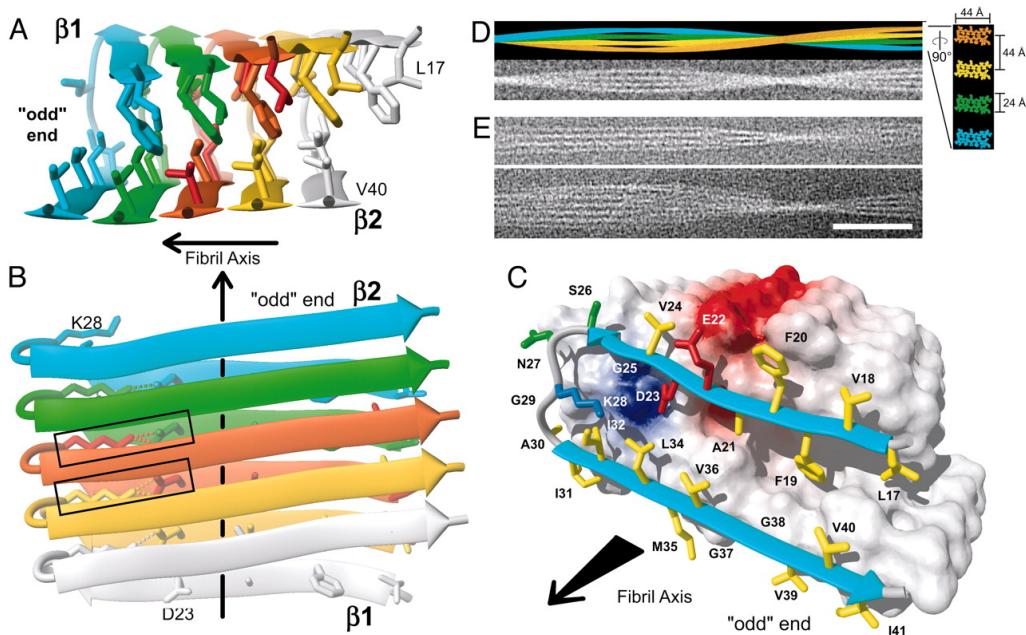


Figura 1.9: **Estructura de fibrillas amiloides.** Figura extraída de [70]

A pesar de esta arquitectura claramente definida de las fibrillas amiloides, las proteínas que forman estos agregados son muy diversas, con muy poca similitud entre sus secuencias. Lo que es más relevante, estudios con diversos péptidos y proteínas no relacionados con aquellos encontrados como causantes de la amiloidosis sistémica, muestran que todos son capaces de formar, bajo las condiciones apropiadas, agregados con las características de las fibrillas amiloides. Estos estudios experimentales, junto con una gran variedad de análisis computacionales, llevan a pensar que la estructura amiloide puede ser adoptada por cualquier cadena polipeptídica y, por lo tanto, esta habilidad para reordenar la conformación monomérica y agregarse para formar la estructura amiloide característica puede ser una propiedad intrínseca de las proteínas [46].

Sin embargo, la tendencia a formar este tipo de fibrillas amiloides sí es un proceso

que depende específicamente de la secuencia. Algunos detalles estructurales a nivel atómico, así como la estabilidad relativa del estado amiloide, pueden ser también características fuertemente dependientes de la secuencia. Por lo tanto, la tendencia a formar estos estados agregados puede variar ampliamente entre distintas proteínas. La pregunta clave es ¿cómo influye la secuencia o composición de aminoácidos sobre esta tendencia?

Se sabe que las interacciones hidrofóbicas son esenciales para la compactación de la estructura lineal de un polipéptido. Se podría esperar, entonces, que un incremento en el contenido de residuos hidrofóbicos pueda aumentar la tendencia a la agregación de una secuencia [56], mientras que un incremento en la carga neta impida esta agregación [49]. Además de estas propiedades fisicoquímicas, dadas las características estructurales recién descritas para las fibrillas amiloides, con un alto contenido de hojas- β , es razonable pensar que las secuencias que tienen una tendencia intrínseca a adoptar conformaciones de hoja- β en su estructura secundaria probablemente tendrán, también, cierta tendencia a la formación de estos agregados.

La composición secuencial parece tener un papel importante en la tendencia a la agregación, sin embargo, la combinación lineal de estas propiedades a lo largo de la estructura primaria puede tener un impacto aún mayor, es decir, no todas las secuencias de polipéptidos con la misma composición tienen la misma tendencia a la agregación [35]. En [109] se demuestra la existencia de ciertos tramos cortos en la secuencia que promueven la agregación en estructuras amiloides. Estos tramos cortos, llamados generalmente regiones propensas a la agregación (o APRs), se distinguen por tener composiciones características de aminoácidos, ricas en residuos hidrofóbicos.

Además del análisis secuencial, existen otras evidencias a favor de la hipótesis de que algunas regiones pequeñas de las proteínas son responsables principales de la formación de fibrillas amiloides. Por ejemplo, la mayoría de las IDPs no experimentan agregación *in-vivo* [67] lo que indica que, si bien la conformación no-plegada es necesaria, no es condición suficiente para promover el proceso de agregación. Por lo tanto, deberían existir patrones secuenciales distinguibles que, una vez expuestos (fuera del núcleo hidrofóbico de la estructura plegada), son más propensos a la agregación que otros.

Los agregados proteícos están generalmente asociados con condiciones anormales donde se pierde la capacidad de mantener las proteínas en sus estados nativos correspondientes y, por lo tanto, estas pierden la capacidad de realizar sus actividades biológicas correctamente, con consecuencias negativas para el sistema. En particular las fibrillas amiloides fueron descubiertas en el contexto de una enfermedad sistémica

y se cree que están asociadas a una gran cantidad de enfermedades, sin embargo, han sido aprovechadas por distintos sistemas biológicos y pueden encontrarse versiones funcionales de estos agregados en distintos organismos [48].

1.2.3.5. Continuo resultante de estructuras, secuencias y actividades

Al estudiar las propiedades conformacionales de las proteínas intentamos clasificar los conocimientos en distintos elementos modulares (proteínas globulares, intrínsecamente desordenadas, etc), lo cual sirve como modelo para entender y predecir las propiedades de interés (conformacionales y funcionales principalmente). Sin embargo, dado que la ocurrencia de proteínas está ajustada a los requerimientos del sistema y al mecanismo evolutivo subyacente, la realidad es que, incluso si tratamos de agruparlas en clases distintas, el perfil de muchas propiedades asociadas a las proteínas muestran un perfil que es mejor descrito mediante un continuo de posibilidades. Ejemplo de esto es el perfil de propiedades conformacionales encontrados en las proteínas. A pesar que intentamos distinguir perfiles conformacionales con estructuras tridimensionales definidas de aquellas que no adoptaban una estructura ordenada, el perfil conformacional real que se observa en el proteoma forma un continuo estructural similar al que se ve en la figura 1.10.



Figura 1.10: Descripción gráfica del continuo conformacional: Los segmentos desordenados pueden comprender solo un pequeño conjunto de aminoácidos dentro de las proteínas u ocupar grandes segmentos de ésta, resultando en un perfil similar al que se ve en la figura. De izquierda a derecha: estructura totalmente ordenada; extremos N y C terminales desordenados; región linker desordenada uniendo dominios globulares; loop desordenado; dominio completamente desordenado; proteína desordenada con elementos estructurales residuales; proteína desordenada sin elementos estructurales definidos pero con conformación colapsada; proteína totalmente desordenada.

Algo similar ocurre con los elementos funcionales emergentes de la secuencia. Por ejemplo, si bien hay diferencias conceptuales entre los elementos que definimos como motivos secuenciales y los MoREs, también existen varias propiedades en común incluyendo la tendencia intrínseca al desorden, ser propensos a experimentar transiciones a estructuras ordenadas y a promover la formación de complejos. Varios estudios sugieren, entonces, que existen un gran solapamiento entre estos elementos que normalmente tratamos como conjuntos definidos y diferentes [50, 76].

Dependiendo si la idea es definida/estudiada desde un punto de vista estructural o secuencial, entonces, un motivo lineal puede ser también identificado como un MoRE. Más allá de esto, a pesar de tener una longitud que parece ser bastante mayor, los IDPs también poseen un solapamiento significativo con MoREs y motivos lineales. Estos resultados parecen indicar que estos elementos funcionales que identificamos con propiedades diferentes pertenecen a estados de un mismo continuo de mecanismos de unión que pueden encontrarse en IDPs.

1.3. Ingeniería de proteínas

1.3.1. Ingeniería de proteínas modulares

1.3.1.1. Conceptos generales

Como producto de los avances logrados en las tecnologías asociadas al ADN recombinante, se ha desarrollado una nueva generación de proteínas compuestas por la integración de diferentes módulos secuenciales.

La idea de armar proteínas a partir de la unión de módulos no es algo nuevo sino que sigue la lógica presentada previamente de que las proteínas naturales son usualmente modulares. Es decir, se está simulando el proceso evolutivo natural que desarrolla nuevas proteínas mediante la combinación de dominios preexistentes.

La utilización de la técnica de ADN recombinante para construir nuevas proteínas abre toda una gama de posibilidades que van desde inserción de pequeñas secuencias en extremos de proteína naturales, con el fin de poder identificarlas o separarlas, hasta el diseño de construcciones proteicas que buscan obtener proteínas con nuevas funcionalidades o propiedades diferentes.

En los primeros casos el proceso experimental suele ser bastante directo y basta con fusionar un segmento en el extremo deseado de la proteína. En los últimos casos de uso, la implementación del proceso experimental tiende a ser más complejo, requiriendo una aproximación de ingeniería. En ésta se pueden distinguir varios aspectos de diseño a considerar, los cuales no siempre son totalmente independientes entre si: por un lado está el proceso de diseño de la nueva proteína, por otro lado están los aspectos asociados con la técnica experimental de ADN recombinante y expresión de proteínas heterólogas. En el gráfico 1.11 se representan los pasos generales que pueden formar parte de este proceso.

El diseño de la proteína de fusión (químérica) requiere de dos elementos indispensables: los dominios o proteínas a fusionar, y la secuencia de enlace (linker) que los va a unir. La elección de los dominios/proteínas está fuertemente ligada al pro-

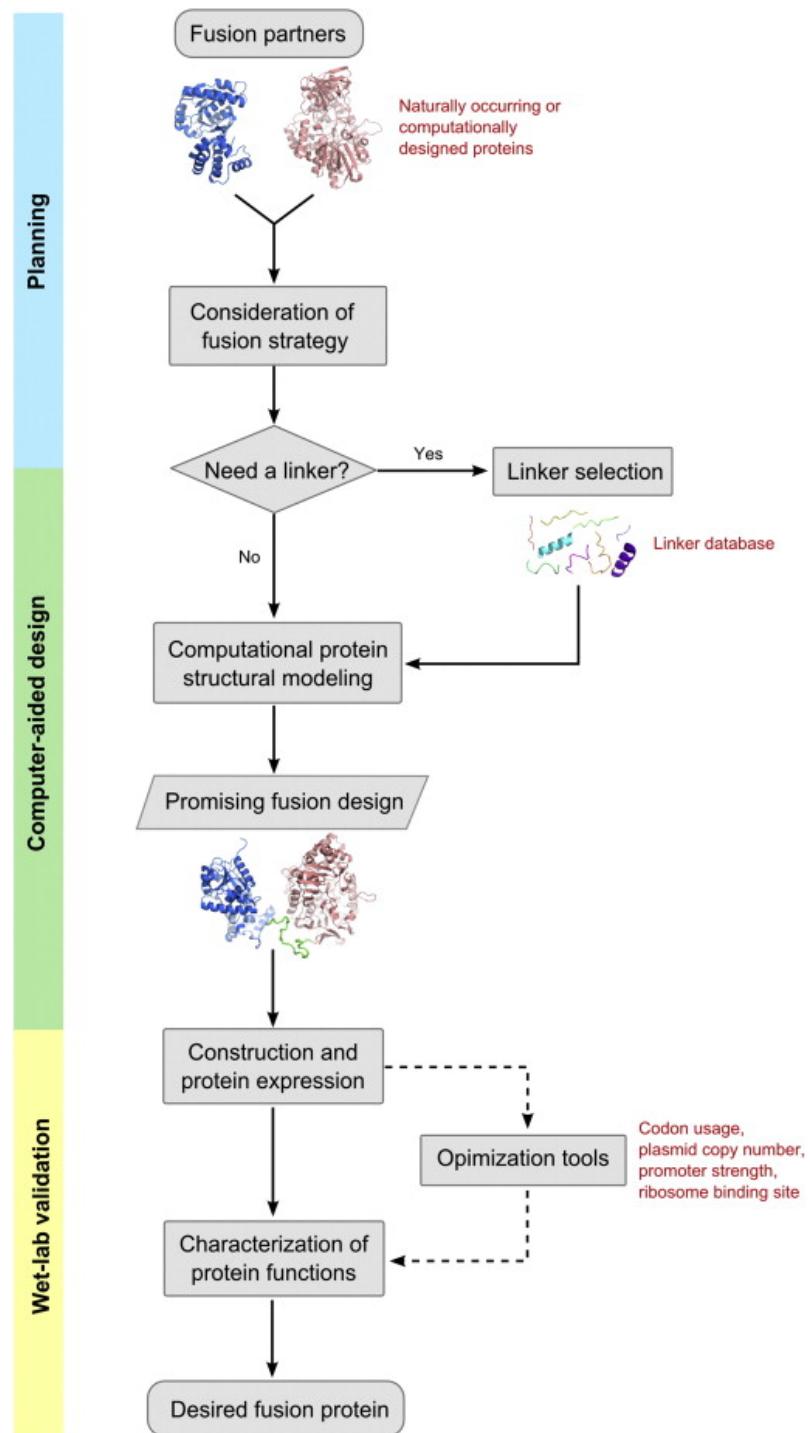


Figura 1.11: Proceso de construcción de proteínas químéricas. Figura obtenida de [118]

ducto final que se desea obtener y, normalmente, es una decisión directa alrededor de la cual se diseña el resto del experimento. Por otro lado, la selección de un linker adecuado para unir los dominios/proteínas de acuerdo al objetivo que se busca puede ser un paso complicado y es generalmente ignorado cuando se piensa en el diseño de una proteína químérica. La unión directa de los dominios/proteínas o el uso de un linker inadecuado puede resultar en resultados indeseables. Por ejemplo, puede restringirse la capacidad de plegado de algún dominio globular, disminuir el rendimiento de la proteína resultante o la actividad biológica de alguno de los módulos. La correcta selección, o mejor aún, el diseño racional de las secuencias linker para unir los módulos es un aspecto importante, aunque poco desarrollado, del diseño de proteínas químéricas. Esta falta de desarrollo en el tema se debe quizás a la falta de conocimiento sobre los factores estructurales que gobiernan la flexibilidad entre los dominios, y resulta ser un factor claramente limitante en el diseño *de-novo* de proteínas químéricas. Los conocimientos relevantes sobre estos aspectos estructurales de las secuencias linker tema pueden estar disponibles actualmente a partir de la gran cantidad de secuencias que se disponen, los avances en la identificación automática de dominios y las nuevas técnicas para predecir propiedades conformacionales a partir de la información secuencial.

1.3.1.2. Ejemplos

La gran cantidad de posibilidades de aplicación que tiene la ingeniería de proteínas químéricas alienta a seguir investigando los aspectos básicos de las proteínas asociados a este proceso y a desarrollar técnicas que ayuden en el proceso de diseño. Algunos ejemplos de aplicación son:

- Creación de enzimas multifuncionales[45, 69]: La forma más común de crear proteínas químéricas es fusionar genéticamente dos o más dominios con funcionalidades distintas, obteniendo una nueva función global para la proteína, resultante de la suma de éstas. En el caso de enzimas el objetivo generalmente es que una misma construcción pueda tener afinidad por distintos sustratos, o que las distintas enzimas formen una cadena metabólica, donde el producto de una sea a la vez el sustrato de otra. Es importante que el linker que une a estas distintas enzimas permita la correcta actividad de cada una por separado.
- Facilitar el estudio de interacciones proteína-proteína[88]: Tradicionalmente, la forma más simple para estudiar la interacción entre dos proteínas es expresarlas conjuntamente para obtener el complejo formado. Sin embargo, si la afinidad es muy baja, no es tan simple que este complejo se forme de manera estable.

La creación de una proteína quimérica que une a los dominios/proteínas que forman el complejo permite mantenerlas unidas covalentemente mediante una secuencia linker. La flexibilidad de esta secuencia debería permitir la interacción libre entre los componentes generando la formación del complejo. Al estar unidas covalentemente, habrá una posibilidad mucho mayor de interacción y aumentará la cinética de formación, permitiendo realizar los estudios biofísicos necesarios sobre éste.

- Construcción de sensores FRET: La técnica FRET se basa en utilizar el efecto de transferencia de energía entre dos cromóforos(un donor y un aceptor). Dado que la eficiencia en la transferencia, y por lo tanto de la señal de salida, depende de la distancia, este efecto puede ser utilizado en la técnica para evaluar la distancia entre dos cromóforos en el transcurso de algun proceso molecular de interés.

La construcción del sensor generalmente implica utilizar un par de dominios de unión cuya interacción estable depende de algun evento molecular de interés, como la unión de algún elementos o la ocurrencia de algún proceso. Cada dominio de unión es unido por separado a uno de los cromóforos(donor o aceptor) y estas construcciones se unen entre si por linkers flexibles que les permiten interaccionar libremente. En su forma libre, esta construcción permite a los cromóforos moverse de manera independiente. Por lo tanto no ocurrirá una transferencia significativa de energía. Cuando los dominios fusionados a cada cromóforo se unén (lo cual depende de la ocurrencia del evento de interés), la distancia quedará fijada y esto cambiará el perfil de transferencia de energía.

Este esquema se muestra gráficamente en la imagen 1.12

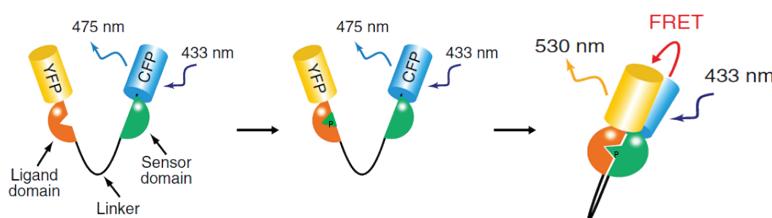


Figura 1.12: **Sensor FRET**. Figura obtenida de [29]

Esta técnica es muy versátil y por lo tanto existen una gran cantidad de ejemplos y modificaciones [87, 110]. En muchos casos la dependencia del proceso molecular con la distancia está directamente relacionada con la secuencia linker. Por ejemplo, si esta secuencia contiene elementos target de modificaciones post-traduccionales que pueden cambiar las propiedades conformacionales del

linker e, indirectamente, el perfil de distancias entre los extremos. En otros casos la conformación en la que los dominios están unidos fija una distancia importante entre los cromóforos. Entonces el sensor mide la disminución en la transferencia de energía como evento resultado del evento de interés. En todos los casos es importante el proceso de ingeniería aplicado sobre el linker para obtener el correcto funcionamiento del sensor.

1.3.2. Ingeniería de secuencias linker

1.3.2.1. Aspectos de diseño

La función de la secuencia linker es unir los dominios covalentemente a la vez que estos actúan como unidades independientes, manteniendo sus capacidades de plegado, funciones individuales y permitiendo que se muevan e interaccionen libremente. Esta funcionalidad depende directamente de la longitud y propiedades conformatacionales del linker, por lo tanto, son los primeros aspectos a tener en cuenta en el diseño de esta.

A pesar de ser un factor directo de la eficiencia[90], la funcionalidad del linker no suele ser tan sensible a la longitud. Un amplio rango de longitudes pueden resultar efectivas, siempre que se eviten las secuencias muy cortas, que no logran obtener una separación suficiente, o las secuencias extremadamente largas, que harían que los dominios funcionen de forma totalmente independiente entre si.

Por su parte, los requerimientos conformatacionales suelen ser más estrictos e incluso pequeños cambios en la estructura pueden limitar la funcionalidad del diseño.

1.3.2.1.1. Propiedades conformatacionales

En primer lugar buscamos que la secuencia linker adopte una conformación desordenada. Es decir, que no se pliegue formando una estructura compacta con los dominios que une. Esta conformación provee la flexibilidad necesaria para que los dominios puedan moverse libremente y las interacciones puedan ocurrir sin ninguna restricción estructural. Para lograr esto es importante controlar que la secuencia no sea propensa a formar estructuras secundarias rígidas tales como α -hélices o hebras- β , ya que esto limitaría la flexibilidad, afectando directamente la interacción entre los dominios que une. Estas propiedades conformatacionales deseadas se corresponden con las características vistas para las proteínas intrínsecamente desordenadas.

En la figura 1.13 se muestra cómo influyen estos requerimientos en la construcción de una proteína químérica que une dos dominios EBFP y EGFP. Los gráficos A,B y E muestran conformaciones globales de la proteína que son posibles gracias a la

flexibilidad del linker intrínsecamente desordenado con una longitud apropiada. Las interacciones que permiten estas conformaciones no se pueden obtener cuando se usa una secuencia con estructura de estructura de hélice- α como se ve en las figuras C y D, donde la rigidez de ésta hace que las conformaciones globales sean mucho más limitadas y ciertas interacciones no puedan ocurrir.

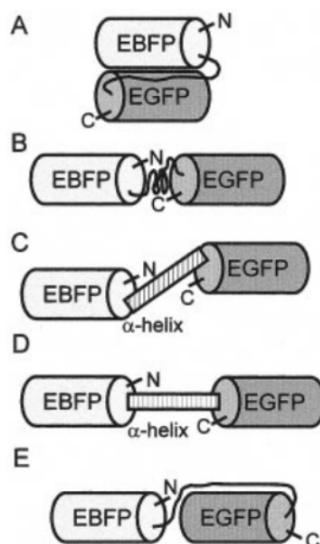


Figura 1.13: **Linkers adoptando diferentes conformaciones.** Figura obtenida de [21]

Las características conformacionales buscadas no sólo serán resultado de interacciones intramoleculares sino que dependen del contexto en el cual la proteína diseñada se encuentre. La construcción química se expresa en un sistema biológico determinado a partir del cual puede ser luego aislada y purificada para estudiar. Esto impone ciertos requerimientos al diseño del linker ya que las interacciones estables con otros componentes del sistema pueden limitar seriamente la flexibilidad intrínseca de este.

Uno de los procesos que pueden afectar más severamente la funcionalidad del linker es la formación de agregados proteicos durante la producción de la proteína[66]. En este caso la unión a otras proteínas no solo limita la flexibilidad del linker sino también arrastra a la proteína completa inhibiendo totalmente la actividad de esta. Además, la agregación no es sólo un problema de la expresión en el sistema biológico, donde la proteína debe mantenerse en solución, sino que también puede ocurrir durante las etapas posteriores de separación y purificación. Es importante que el diseño del linker tenga en cuenta estos aspectos, evitando que la secuencia posea tendencias a formar uniones estables con otros componentes del sistema.

1.3.2.1.2. Elementos biológicamente funcionales

Los aspectos conformacionales no son los únicos relevantes. Es importante que, además, la secuencia se mantenga biológicamente inerte frente al sistema en el cual se expresa de manera que el linker funcione únicamente como conector entre los dominios. Incluso cuando se trate de secuencias con longitud corta, es importante que no posea ningún elemento secuencial que pueda funcionar como módulo de interacción iniciando un proceso biológico.

La gama de actividades que involucra a los módulos funcionales, aún aquellos que puedan encontrarse en secuencias cortas (como suelen ser los linkers), es muy diversa. Por lo tanto, pueden tener implicancias diferentes tanto sobre el linker y la proteína químérica como en el normal funcionamiento del sistema biológico donde se expresa esta. Ejemplo de esto son los sitios target de clivaje, cuya ocurrencia dentro de la secuencia linker tendría consecuencias obvias que afectarían directamente a la unión covalente que genera el linker. Otros ejemplos son las modificaciones post-traduccionales sobre la secuencia, las cuales pueden modificar las características de esta (por ejemplo la carga neta) impactando directamente en las propiedades conformacionales asociadas.

Muchas veces un mismo proceso representa efectos biológicos y conformacionales negativos conjuntamente, tal es el caso de los MoREs, los cuales usualmente señalan el desarrollo de un proceso biológico mediante un mecanismo conocido como plegamiento y unión, en el cual se estabilizan estructuras secundarias donde antes existía una conformación desordenada.

1.3.2.1.3. Otras propiedades

Como parte del proceso de producción se requiere que el sistema exprese la construcción diseñada, lo cual, al ser una proteína foránea, estará directamente ligado a la carga metabólica que la secuencia imponga sobre el sistema [52]. Esta carga metabólica es el resultado de la utilización de recursos propios de la célula para la expresión de esta nueva proteína y depende, entre otros aspectos, de la composición de aminoácidos de esta. Diferentes aminoácidos representan diferentes requerimientos metabólicos, por lo tanto, un aspecto relevante de la secuencia es que esta tenga una composición que imponga mínimos requerimientos metabólicos al sistema. La composición raramente puede ser modificada en los módulos que se están fusionando, pero sí es posible tener estos aspectos en cuenta al momento de diseñar la secuencia linker. La importancia de este requerimientos depende de las consideraciones del usuario que diseña el experimento, del sistema de expresión que se va a usar, el nivel de expresión requerida, etc.

En otros casos, los pasos posteriores a la expresión pueden imponer requerimientos sobre el linker que provienen de la manipulación, detección o procesamiento de la proteína construida. Por ejemplo, dado que las técnicas espectroscópicas se usan de forma rutinaria en el trabajo con proteínas, es posible que el usuario quiera realizar la construcción utilizando un linker cuya composición específica no posea aminoácidos con absorbancia en el rango del UV, de forma que se elimine cualquier interferencia con este tipo de técnicas.

Todas estas situaciones están asociadas a la composición del linker diseñado. La composición ideal de esta secuencia linker dependerá, entonces, no sólo de las propiedades secuenciales asociadas a la funcionalidad dentro de la construcción, sino también de aquellos aspectos que el usuario considere relevantes como parte del proceso experimental.

1.3.2.2. Linkers naturales

1.3.2.2.1. Características

La modularidad en proteínas naturales es algo muy común y existen muchos ejemplos de proteínas compuestas por dos o más dominios funcionales unidos mediante linkers. Estas secuencias linker sirven, principalmente, para mantener unidos los distintos dominios y proveer la flexibilidad necesaria para que puedan ocurrir las interacciones asociadas a la actividad biológica de la proteína (ver figura 1.14). Esta es la funcionalidad básica inherente a todos los linkers naturales, y está directamente asociada a la dinámica conformacional de cada proteína.

Los estudios realizados sobre secuencias linkers naturales [22, 27, 51] muestran que, en general, los residuos encontrados en estas pertenecen al grupo de aminoácidos polares (con o sin carga), y son específicamente ricos en Prolina. La ocurrencia de Prolina es bastante esperable, puesto que su cadena lateral cíclica y la falta de un hidrógeno disponible en su grupo amida previene la formación de puentes de hidrógeno con otros aminoácidos, dándole una baja tendencia a formar estructuras secundarias típicas (hélices- α y hojas- β). Por lo tanto, las secuencias ricas en prolina resultan en una cadena carbonada con estructura más bien rígida. Como parte de un linker, entonces, la presencia de prolina permitiría prevenir la formación de una estructura plegada con interacciones desfavorables entre los dominios.

Todas las propiedades analizadas (longitud, composición, hidrofobicidad y estructura secundaria) resultaron importantes para alcanzar las funciones deseadas. En general, se encontró que los linkers naturales adoptaban principalmente conformaciones desplegadas y tenían estructuras independientes sin interacciones con los dominios adyacentes. En términos de estructura, se encontraron tanto linkers flexibles como

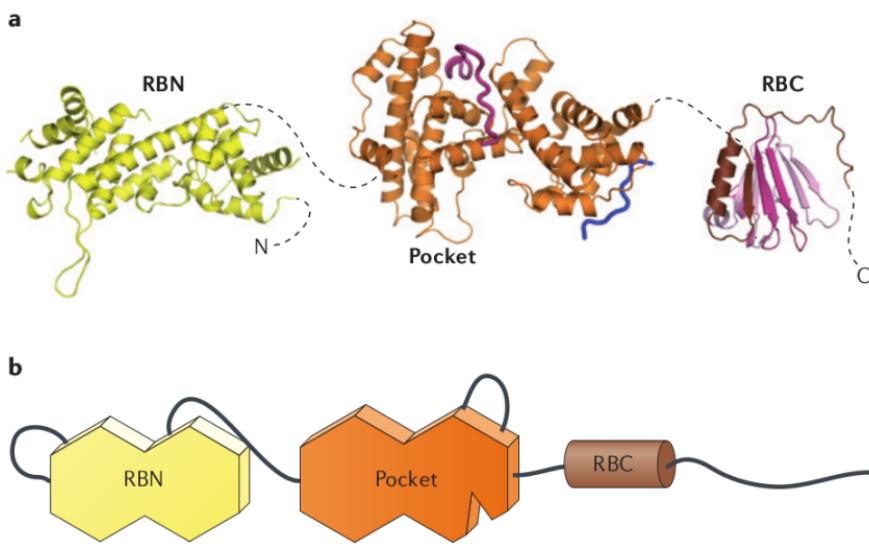


Figura 1.14: **Ejemplo de linkers naturales de la proteína RB (Retinoblastoma).** Los linkers y loops se muestran en a mediante lineas entrecortadas. En b se muestra una representación más abstracta de la misma estructura multidomínio. Figura extraída de [37]

relativamente rígidos, estos últimos formados principalmente por α -hélices en su estructura secundaria. Basándose en la frecuencia de este tipo de estructuras, en [51] se esboza una clasificación de los linkers dividiéndolos en dos categorías: helicoidales y no helicoidales.

En términos de flexibilidad, y por lo tanto de propiedades conformacionales en general, los linkers naturales muestran un rango muy amplio de características, las cuales se pueden apreciar mejor a partir de algunos ejemplos. En algunos casos la flexibilidad provista se origina a nivel de estructura terciaria, y está limitada a una región corta que funciona como bisagra. En otros casos la flexibilidad es “completa” y la región linker es intrínsecamente desordenada[71], encontrándose incluso que la funcionalidad de la proteína se pierde cuando se reemplaza por un linker que adopta una conformación estructurada como en el caso de una hélice- α [57]. Además existen linkers que, a pesar de mantener una conformación extendida en solución, pueden plegarse (o fijar una estructura secundaria transiente) en presencia de ligandos. Tal es el caso de ciertas proteínas que se unen a ADN[65]. En otras proteínas, las propiedades conformacionales de la secuencia linker sirven para mediar la propagación eficiente de estímulos entre los dominios que conectan (alosterismo). Ejemplo de esto es la miosina del músculo liso, en la cual el dominio que comprende la función motora puede ser activado mediante fosforilación de un dominio regulatorio, el cual

se encuentra conectado mediante una región linker[59].

Por lo tanto, a pesar que la flexibilidad es un aspecto que se sabe esta ligado a las secuencias linker en la naturaleza[116], y que estas regiones permiten los grandes cambios conformacionales en las estructuras de las proteínas, no son simplemente secuencias que evolucionan hacia regiones totalmente flexibles, ya que cada proteína impone requerimientos conformacionales específicos.

De esta forma, podemos ver a las secuencias linker como un módulo más de las proteínas naturales, cuyas propiedades conformacionales no pueden definirse en forma generalizada, sino que están asociadas a restricciones funcionales de la proteína, buscando balancear la flexibilidad requerida para que los dominios tengan la dinámica que su funcionalidad requiere, con la rigidez necesaria para que no existan interacciones desfavorables entre estos, resultando en perfiles estructurales muy variados.

1.3.2.2.2. Utilización en diseños de proteínas quiméricas

Si bien todos los linkers naturales permiten mantener covalentemente unidos a los dominios, las propiedades conformacionales pueden ser muy distintas a las que buscamos para un linker en el contexto de una proteína químérica. Por otro lado, en cada proteína natural, la secuencia linker puede proveer funciones que forman parte directa de la actividad biológica correspondiente, tales como intervenir en las interacciones cooperativas entre los dominios, contener elementos secuenciales que median interacciones con otras proteínas o ligandos, etc. Por lo tanto, al utilizar un linker natural en una proteína químérica podríamos estar agregando un módulo biológicamente funcional a la proteína que estamos diseñando.

Dado este contexto, el uso de secuencias linkers naturales en el diseño de proteínas químéricas debe realizarse con mucho cuidado ya que no todos cumplen con los requerimientos conformacionales y funcionales buscados. Mas allá de esto, aún en el caso de encontrar secuencias naturales que cumplan los requerimientos buscados, la diversidad de secuencias existente, principalmente en cuanto a composición y longitud, puede no ajustarse a la diversidad de requerimientos que se presentan a partir de la utilización experimental de proteínas químéricas.

Poder usar secuencias naturales presenta, sin embargo, una ventaja fundamental. Generalmente se han estudiado en detalle las propiedades conformacionales y funcionales asociadas a los linkers naturales en el contexto de sistemas biológicos. De otra forma, existe información disponible para poder inferir sus propiedades a partir de análisis secuenciales comparativos. En cualquier caso, podemos tener una idea aproximada de qué esperar si utilizamos estas secuencias como módulo de una

nueva proteína expresándola en un sistema biológico real. Por último, más allá de las propiedades finales de la proteína, las secuencias nucleotídicas que codifican linkers naturales han evolucionado como parte de sistemas biológicos reales, de forma que no suelen plantear ningún desafío para la expresión.

1.3.2.3. Diseño racional

1.3.2.3.1. Diseños y conceptos comunes

Con tantos y tan distintos requerimientos para el diseño, el problema de encontrar un linker adecuado puede llegar a ser un proceso complejo. Los diseños encontrados en la literatura están generalmente basados en la intuición y en propiedades generales de los linkers naturales. Además, generalmente implican un proceso iterativo de evaluación y optimización experimental del diseño base. En [27] se revisan los diseños resultantes de este proceso artesanal que han sido finalmente utilizados en distintos contextos experimentales.

Basándose en los aspectos conformacionales de interés, un linker poli-G parece ser la opción evidente para proveer a la construcción de máxima flexibilidad. Sin embargo, como vimos antes, la flexibilidad no es todo. Por ejemplo, una cadena poli-G puede resultar en una conformación poco soluble debido a que no contiene ningún residuo polar, afectando la actividad de los dominios que une[90]. Por otro lado, la secuencia nucleotídica que codifica a este polipéptido tiene un alto contenido de Guanina (los codones que codifican para Glicina contienen Guanina en 2 de las 3 posiciones), lo cual puede hacerla difícil de manejar experimentalmente y de expresar en el huésped[102]. De esta forma, la secuencia poli-G puede no ser favorable para utilizar como linker aunque se ha utilizado en algunos casos[32, 60, 93], y es la base de gran parte de los linkers flexibles encontrados, sobre los cuales se agregan algunas modificaciones puntuales usando otros aminoácidos pequeños como Serina.

El diseño de linker más común encontrado en la literatura está compuesto por un número variable de repeticiones del motivo (G_4S). En estos casos la sustitución de algunas posiciones por residuos polares (Ser) busca solubilizar la secuencia poli-G, permitiendo interacciones entre la cadena lateral polar y el medio acuoso. Al usar esta secuencia como linker, esto previene interacciones no deseadas con los dominios que une, de forma tal que no interfiera en la funcionalidad. Una aproximación de diseño más avanzada se puede ver en [24], donde se utiliza un linker natural como diseño base y se agregan residuos de Glicina y Serina para proveer flexibilidad, y residuos de ácido Glutámico y Lisina para incrementar la solubilidad. La secuencia resultante es EGKSSGSGSESKST.

En todos estos casos, los péptidos con motivos repetidos de G y S son usados

porque se asume que adoptan una conformación similar a random coil y no interfieren en el plegado y funcionamiento de los dominios que unen. Sin embargo, en cada caso de aplicación particular se debe evaluar experimentalmente que el diseño cumple el objetivo buscado. Estos linkers poseen conformaciones desestructuradas, dado que la glicina es capaz de romper la estructura ordenada de las hélices. Esta flexibilidad puede, en algunos casos, impedir que se logre una separación suficiente entre los dominios [44]. En este último caso, donde se requiere cierta flexibilidad pero al mismo tiempo mantener suficiente distancia entre los dominios, se han utilizado linkers que adquieren una estructura completa de α -hélice. El ejemplo más común de linkers con estas propiedades es $(EAAAK)_n$ [20]. En otros casos, la rigidez se puede obtener utilizando secuencias poli-P [95]

Como se ve, si bien hay un patrón en los diseños, que contienen prioritariamente residuos G y P, en cada caso se analizan los requerimientos específicos del nuevo diseño, definiendo individualmente los residuos en cada posición de la secuencia.

1.3.2.3.2. Algoritmos existentes

Los ejemplos de diseño descritos muestran cierta sistematización y racionalidad en el proceso utilizado agregando distintos residuos según las necesidades. Sin embargo, la creación de un método de diseño racional para secuencias linkers basado en los requerimientos está aún en los principios del desarrollo. Los estudios detallados de la composición, estructura y función de linkers naturales son un claro punto de inicio, abriendo la posibilidad de crear bases de datos conteniendo las secuencias encontradas y sus propiedades asociadas. Asociados a estas bases de datos, se desarrollado distintos métodos de búsqueda que permiten extraer secuencias con propiedades deseadas específicas, lo que constituye un mecanismo simple de diseño.

Los resultados del estudio desarrollado en [51] son el primer ejemplo de este tipo de metodologías. En este trabajo se estudian diversos aspectos de secuencias linkers, las cuales son extraídas mediante un método automatizado a partir de una base de datos de estructuras proteicas. A partir de las secuencias recolectadas se desarrolla una base de datos asociada a un algoritmo de búsqueda, el cual puede ser utilizado mediante un servidor web[58]. El método de búsqueda implementado acepta distintos parámetros como longitud del linker, accesibilidad del solvente, estructura secundaria, similitud con una secuencia input, etc. El programa devuelve las secuencias que contienen los criterios solicitados y, además, provee información del contexto en el que se encuentra en la proteína natural, con información del identificador PDB, descripciones funcionales de la proteína, etc., de forma que el usuario pueda inferir otras propiedades del linker que no son presentadas automáticamente.

Un ejemplo más reciente de este tipo de aproximación mediante bases de datos da origen a la herramienta LINKER [28, 117]. Esta aplicación posee un método de búsqueda que brinda una gran cantidad de opciones al usuario, incluyendo aspectos experimentales como la sensibilidad a la actividad de proteasas. En este caso, el centro del método es una base de datos conteniendo secuencias bucle extraídas del PDB y que son luego levemente procesadas retirando secuencias idénticas, bucles de horquillas beta (hairpin loops) y secuencias de menos de 4 residuos. El programa de búsqueda sobre esta base de datos asume que la conformación de loop adoptada por la estructura cristalizada encontrada en la PDB dará siempre una conformación desordenada al utilizarse como linker en una proteína químérica. Este es un aspecto que debe ser comprobado experimentalmente para cada caso ya que, como vimos, la flexibilidad puede no estar asegurada. Desafortunadamente, el servidor web asociado a este programa no se encuentra más disponible.

SynLinker[68] representa la versión más actual y completa de este tipo de soluciones para el diseño de linkers. Este nuevo programa incorpora en su base de datos no sólo secuencias linker naturales, sino también algunos linkers empíricos extraídos de la literatura, los cuales siguen los principios de diseño que vistos previamente. Estos últimos, si bien representan sólo una pequeña parte de las secuencias disponibles en la base de datos, tienen la particularidad de haber sido evaluados experimentalmente como linkers en proteínas químéricas lo que les asegura ciertas propiedades particulares que los distinguen de los linkers naturales. En la figura 1.15 se muestran de manera gráfica todas las secuencias que se integran en la base de datos sobre la cual corre el método de búsqueda de SynLinker.

Esta herramienta va más allá en el proceso de diseño visto previamente (figura 1.11), permitiendo obtener un modelo computacional asociado a la construcción que estamos armando. El modelo contiene uno o más linkers resultantes de la búsqueda, unidos a estructuras de dominios seleccionados por el usuario. El mismo modelo puede ser usado directamente como input en el próximo paso (siguiendo el esquema de la figura 1.11) para evaluar mediante simulaciones de dinámica molecular algunas propiedades conformacionales de la construcción obtenida.

En todos estos ejemplos de búsquedas parametrizadas, el proceso de diseño asume que la secuencia obtenida se comportará de la misma forma en la proteína natural que en una nueva construcción, donde se encuentra uniendo dominios completamente diferentes, lo cual debe ser comprobado siempre de manera experimental. Por otro lado, la selección está siempre basada en propiedades conformacionales aunque, como vimos antes, estos no son los únicos requerimientos de las secuencias linker. De esta forma, a pesar que no proveen una solución completa al problema de diseño,

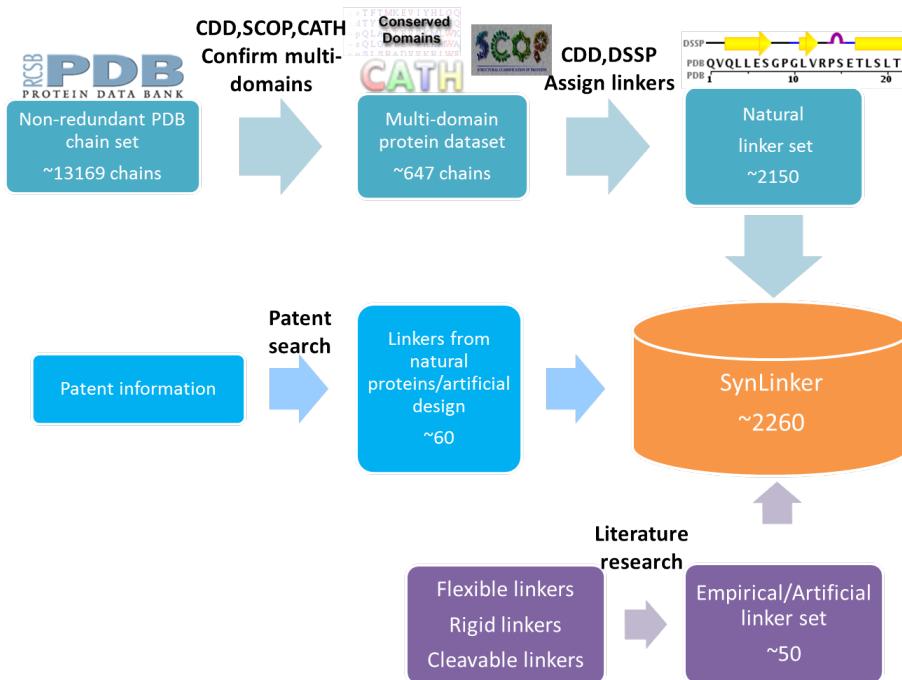


Figura 1.15: **Esquema gráfico de la construcción de SynLinker.** Figura extraída de [68]

la construcción de estas bases de datos y los métodos de búsqueda asociados ayuda a la utilización de los conocimientos adquiridos a partir de estudios sobre secuencias naturales y sirve como primer paso para un proceso posterior de ingeniería, el cual debe involucrar necesariamente la evaluación experimental de la construcción diseñada. El desarrollo de métodos de búsqueda más abarcativos junto con nuevos estudios para encontrar y refinar las secuencias linker naturales podría generar un avance en este tipo de metodologías.

1.4. Objetivos

El objetivo principal de este trabajo es implementar un método que permita generar una secuencia linker *de novo* a partir de los requerimientos descritos, o adaptando una secuencia inicial provista por el usuario. Como requerimientos fundamentales de las secuencias resultantes del método, deberá cumplirse que las mismas provean la flexibilidad necesaria para cumplir su función como linker en el diseño de nuevas proteínas químéricas. Además, deberán mantenerse inertes frente a cualquier actividad que pueda interferir con el proceso experimental asociado o con el correcto funcionamiento del diseño final.

El esquema de la herramienta resultante se muestra, de forma simplificada, en la figura 1.16

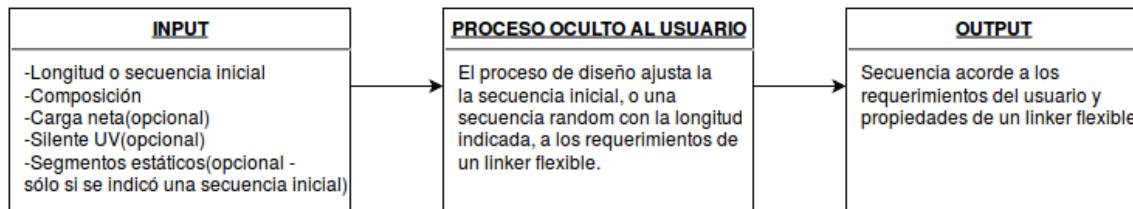


Figura 1.16: Esquema general de la herramienta implementada

Capítulo 2

Herramienta desarrollada

2.1. Fundamentos del método utilizado

2.1.1. Detección de propiedades a partir del análisis secuencial

En la sección 1.3.2.1 se desarrollaron distintos aspectos estructurales y funcionales que se deben tener en cuenta en el diseño de un linker. A partir de esta descripción podemos distinguir diferentes características que afectan positiva y negativamente a la funcionalidad del diseño resultante. Esto permite definir una secuencia linker como aquella que posea todas las características positivas y no posea ninguna de las características negativas buscadas. Es importante recordar que, como se describió previamente, algunas de las características están en función de consideraciones hechas por el usuario experimental.

En la sección 1.2.3 describimos distintos módulos proteicos encontrados en la naturaleza, identificando los elementos conformacionales y funcionales asociados. El estudio de estos elementos generalmente apunta, no sólo a modelizar sus propiedades asociadas (mecanismos funcionales, estructuras, propiedades secuenciales asociadas, etc.), sino también a poder trasladar los conocimientos a herramientas bioinformáticas que permitan predecir la ocurrencia de estos. Estas propiedades están directamente asociadas a la secuencia de la proteína. Además, la secuencia es lo primero que se conoce cuando se encuentra o diseña una nueva proteína, por lo tanto, como objetivo general se intenta poder obtener la mayor cantidad de información a partir de ésta. Como consecuencia, existen cada vez más herramientas, que utilizando aproximaciones distintas permiten detectar diversas características a partir de la secuencia.

Una parte fundamental del método implementado en este trabajo es que actual-

mente podemos predecir, con cierta confianza, la ocurrencia de los distintos aspectos positivos y negativos asociados a un linker conociendo solamente la secuencia de aminoácidos de una proteína. Lo que es más, las predicciones pueden ser mapeadas a posiciones puntuales dentro de la secuencia. Por ejemplo, podemos localizar dentro de la secuencia cuales serán las posiciones que favoreceran la adopción de una conformación intrínsecamente desordenada, lo cual es una característica deseable para el diseño de un linker.

2.1.2. Espacio de secuencias buscadas

De todos los requerimientos de diseño vistos en la sección 1.3.2.1 el más restrictivo es, en principio, que el diseño resultante tenga la flexibilidad necesaria. Esto implica que la secuencia se ajuste a las propiedades de una proteína intrínsecamente desordenada, las cuales identificamos previamente como un conjunto de secuencias claramente sesgadas en su composición, con baja complejidad y generalmente con baja abundancia de varios tipos de aminoácidos. Esto parece indicar que el espacio de secuencias comprendido por éstas debe ser significativamente pequeño con respecto al espacio total de posibles soluciones(todas las posibles combinaciones de aminoácidos), y por lo tanto obtener un diseño que encaje dentro de este espacio deja pocas opciones y hace difícil la búsqueda de la secuencia resultante.

Sin embargo, la realidad es más compleja de lo que parece. Las IDPs no requieren que su secuencia posea un código definiendo el plegamiento y la estructura nativa final. Además, vimos que las IDPs se caracterizan por tener secuencias conteniendo múltiples elementos estructurales y funcionales relativamente cortos, formando un mosaico. De esta forma, el espacio abarcado por las secuencias que poseen las propiedades conformacionales que buscamos, puede no ser tan restringido como se pensabamos y, de hecho, puede tener un tamaño considerable con respecto al espacio total de posibles secuencias. Dentro de este espacio, el resto de las propiedades positivas que buscamos y de las propiedades negativas que queremos evitar, no restringe demasiado el número de soluciones ya que, como vimos, probablemente estén limitadas a segmentos muy cortos sobre la secuencia.

Asumimos, entonces, que el conjunto de diseños posibles, constituido por todas las secuencias que poseen las propiedades positivas y no poseen las negativas, tiene un tamaño considerable con respecto al espacio de posibles soluciones compuesto por todas las combinaciones de aminoácidos. Este conjunto, sin embargo, puede reducirse considerablemente si se tienen en cuenta algunos aspectos definidos por el usuario como una composición muy específica, una carga neta elevada, o una combinación de cualquiera de éstos.

2.2. Esquema general de la implementación

A partir de los aspectos detallados en los fundamentos, una primera aproximación obvia sería realizar una búsqueda sistemática dentro del espacio de secuencias hasta obtener una que resulte favorable de acuerdo a nuestros parámetros de evaluación. El problema radica en que es una forma ineficiente de búsqueda, tiene un alto costo computacional y el resultado(asumiendo que lo encontramos) será una composición al azar. Una aproximación más adecuada, sería utilizar todo el conocimiento que tenemos para poder evaluar la secuencia y usar esto a nuestro favor para guiar la búsqueda. Al guiar la búsqueda utilizando los resultados de las evaluaciones, esta sera mucho más efectiva y el costo computacional sera mucho menor. Además, según el punto de inicio que utilicemos para la búsqueda, podremos intentar obtener secuencias resultantes con cierta similitud a la secuencia inicial.

El procedimiento general del algoritmo consiste en aplicar mutaciones puntuales iterativamente a partir de una secuencia inicial, en busca de una secuencia final con las características deseadas. En cada iteración se proponen y analizan posibles mutaciones de acuerdo a un sistema de evaluación de secuencias. La selección de posibles posiciones a mutar se hace de acuerdo a un muestreo ponderado en función de los resultados de la evaluación secuencial. Además, la aceptación de cada mutación está asociada a un método heurístico de decisión. De esta forma, la secuencia de mutaciones aceptadas para una determinada secuencia inicial no necesariamente será siempre la misma. La búsqueda será guiada por el análisis realizado sobre las secuencias pero no quedará completamente determinada por éste.

En la figura 2.1 se ve un esquema general del método completo. En la sección 2.2.1 se detalla como se obtiene la secuencia inicial y luego en las secciones 4.2.4 y 2.2.3 se describen y ejemplifican los pasos para evaluar la secuencia y obtener una mutación puntual en cada iteración.

El resultado principal de la ejecución es una secuencia que cumple con todas las características positivas que se indicaron y no posee ninguno de los aspectos negativos. En un archivo asociado se refleja el proceso de mutación indicando, para cada paso, cual es la posición mutada y la secuencia resultante. Otras opciones de salida son posibles. En la sección 4.2.6 del manual se describen opciones que permiten, entre otras cosas, realizar una ejecución paso a paso con detalles de cada evaluación realizada, o limitar la salida mostrando únicamente la secuencia resultante.

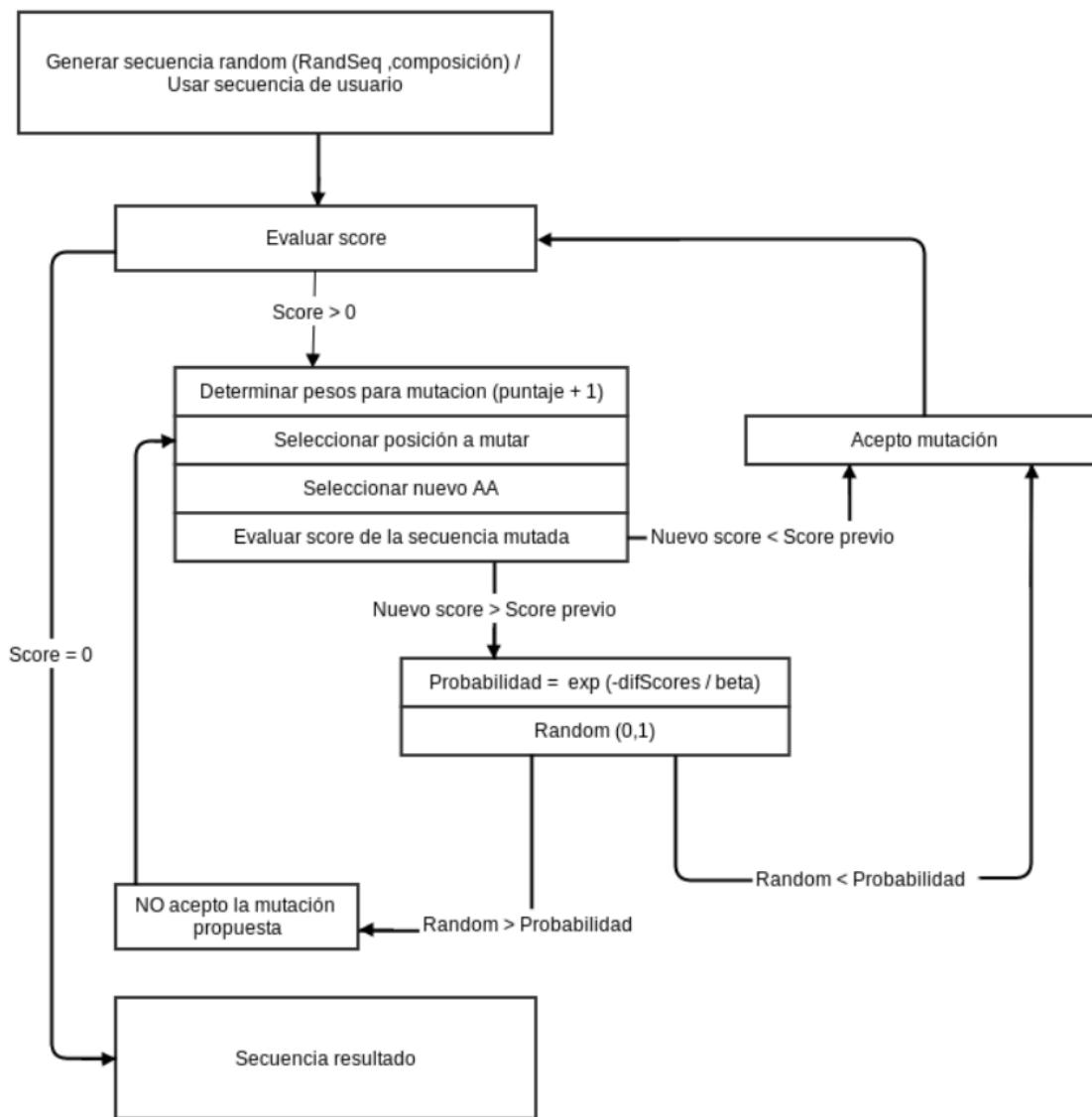


Figura 2.1: Esquema general del método

2.2.1. Secuencia inicial

El método comienza a iterar a partir de una secuencia inicial. Esta secuencia puede ser creada de forma aleatoria como primer paso del algoritmo (ver 4.2.1.1) o puede ser pasada como parámetro por el usuario 4.2.1.2.

En el caso que se defina una secuencia inicial esta definirá también, de manera implícita, la longitud de la secuencia resultante. Por lo tanto, no es necesario definir la longitud. Sin embargo, en este caso se da la posibilidad al usuario de identificar segmentos flanqueantes en uno o ambos extremos de esta secuencia inicial que, si bien serán tenidos en cuenta a la hora de la evaluación, no serán mutados como parte del proceso y por lo tanto formarán parte del diseño resultante. La utilización

de este parámetro opcional se detalla en la sección 4.2.1.3.

La generación de una secuencia aleatoria no es un problema trivial. En la naturaleza las proteínas están compuestas de un conjunto de 20 aminoácidos, los cuales se encuentran con diferentes abundancias relativas. La frecuencia de cada aminoácido dentro de un proteoma está dada por un balance entre el costo metabólico de este y la necesidad de contar con un conjunto de secuencias diversas que darán proteínas funcionales [62]. Siguiendo los mismos principios, definiremos estas frecuencias como la composición estándar para nuestra herramienta. De esta forma, podremos obtener la diversidad buscada a la vez que minimizamos el costo metabólico de las secuencias linker generadas. Las frecuencias estándar resultan del cálculo de la composición de aminoácidos de todas las proteínas de la base de datos SwissProt [1].

Una vez definida la frecuencia que tendrá cada aminoácido, la aplicación que se utiliza para generar una secuencia aleatoria es RandSeq [11]. Entre otras que existen, esta herramienta ofrece la posibilidad de definir las frecuencias que desea para cada aminoácido e, incluso, indicar solo la frecuencia de algunos residuos en particular, dejando el resto con la frecuencia estándar. Nuestra herramienta también ofrece al usuario esta flexibilidad para definir la composición que tendrá la secuencia (ver 4.2.2) Esta funcionalidad permitiría, por ejemplo, indicar que cierto aminoácido no esté presente en la secuencia (asignándole una frecuencia igual a 0). Para el fin que tiene la herramienta, es útil tener este tipo de funcionalidades ya que permite adaptar los requerimientos a las capacidades(limitaciones) del laboratorio experimental. El nuevo linker diseñado deberá poder ser sintetizado eficientemente junto con la nueva proteína, lo cual implica una carga metabólica para el sistema biológico en el cual está siendo expresado. De esta forma, se intentará adaptar las propiedades de la secuencia diseñada para aumentar la capacidad de síntesis reduciendo, por ejemplo, los aminoácidos que implican un gran gasto energético y que limitarán la producción de la proteína final.

2.2.2. Evaluación

En cada iteración del método se analiza la secuencia utilizando un conjunto de herramientas de evaluación. Los resultados de cada evaluación serán reflejados en un valor numérico o puntuación asociado a cada posición (el capítulo 3 se centra en las propiedades evaluadas y cómo se aplican las herramientas utilizadas para dar un valor numérico a cada posición).

El puntaje es inicializado en 0 para todas las posiciones y cada evaluación ejecutada puede aumentar o mantener el valor de manera tal que, al finalizar todas las evaluaciones correspondientes, el puntaje de cada posición tendrá un valor mayor o

igual a cero. En cada evaluación se **aumenta el puntaje si el residuo en esa posición NO favorece las propiedades deseadas**. Definimos el puntaje total de la secuencia como la suma de puntajes de cada posición. Un ejemplo de análisis de la secuencia mediante dos herramientas de evaluación hipotéticas A y B sería:

Secuencia	M	V	L	S	P	A	D	K	T	N	P	D
Puntaje Inicial	0	0	0	0	0	0	0	0	0	0	0	0
Evaluación con herramienta A	0	1	1	1	1	0	3	2	5	4	1	0
Evaluación con herramienta B	0	0	1	1	3	5	1	1	0	2	1	2
Puntaje global	0	0	2	2	4	5	4	3	5	6	2	2
Puntaje total	35											

2.2.3. Mutación

Al terminar todas las evaluaciones sobre la secuencia, el puntaje resultante se utiliza para el proceso de mutación.

El proceso de mutación sigue los siguientes pasos:

1. En primer lugar se selecciona uno de los residuos de la secuencia como objetivo para realizar la mutación. Esta selección será ponderada. El factor de ponderación utilizado será el valor resultante de sumar 1 al puntaje global asociado a la posición. Siguiendo el ejemplo anterior:

Secuencia	M	V	L	S	P	A	D	K	T	N	P	D
Puntaje global	0	0	2	2	4	5	4	3	5	6	2	2
Factor de ponderación	1	1	3	3	5	6	5	4	6	7	3	3

Esta forma de calcular el factor de ponderación implica que nunca se tendrán factores iguales a cero y, por lo tanto, siempre se podrá seleccionar alguna posición para mutar.

2. El segundo paso, consiste en seleccionar con qué aminoácido se sustituirá la posición seleccionada. Esta selección se realiza siguiendo la misma composición que se definió al iniciar la ejecución 4.2.2. Dado que la selección del sustituto es independiente del tipo de residuo seleccionado para mutar, es posible que el mutado y su reemplazo seleccionado sean iguales. En este caso simplemente se vuelve a seleccionar un sustituto hasta que el resultado sea un residuo distinto al anterior.

3. El tercer paso consiste en aceptar o rechazar la mutación propuesta. Para realizar esto, primero se vuelve a evaluar el puntaje, analizando todas las características deseadas de la secuencia pero, esta vez, sobre la secuencia que contiene la mutación propuesta. Este resultado nos permite saber como se modificarían los puntajes si aceptáramos la mutación. La decisión se toma en base al puntaje total de la secuencia, resultante de sumar los puntajes de cada posición. Si este puntaje total disminuye, entonces la mutación es aceptada. Por ejemplo, si evaluamos la posibilidad de mutar la Leucina en la posición 3 por Fenilalanina:

Secuencia previa	M	V	L	S	P	T	D	K	T	N	P	D
Puntaje global	0	0	2	2	4	5	4	3	5	6	2	2
Puntaje total(previo)	35											
Secuencia mutada	M	V	F	S	P	T	D	K	T	N	P	D
Puntaje global	0	0	3	1	3	4	4	3	5	6	2	2
Puntaje total(posterior)	33											

En este caso, la mutación disminuye el puntaje total($33 < 35$) y por lo tanto es aceptada, finalizando la iteración. Una mutación puntual puede cambiar el valor del puntaje en esa posición y/o el valor del puntaje correspondiente a otras posiciones. Por lo tanto, como se ve en el ejemplo, para evaluar el cambio se tiene en cuenta el puntaje total y no exclusivamente el de la posición mutada.

Si la mutación propuesta **NO** disminuye el puntaje total de la secuencia, entonces la mutación se acepta o rechaza siguiendo el método de decisión de Monte Carlo.

Este método deriva una probabilidad de aceptación($P_{aceptar}$) a partir de la diferencia entre los puntajes totales resultantes de la secuencia antes($Puntaje_{previo}$) y después($Puntaje_{posterior}$) de la mutación propuesta. La probabilidad de aceptación resulta de calcular:

$$P_{aceptar} = e^{\left(\frac{Puntaje_{previo}-Puntaje_{posterior}}{\beta}\right)} \quad (2.1)$$

El valor del parámetro β es propio del método y permite ajustar la probabilidad de aceptación según el perfil de evaluaciones realizadas. En el capítulo 5 se analizan en detalle los efectos globales de este parámetro sobre la herramienta y se evalúan los valores óptimos en función del conjunto de herramientas

aplicadas(detalladas en el capítulo 3). Para ejemplificar el método de decisión de Monte Carlo, tomemos el siguiente caso de mutación en donde se propone mutar el ácido aspártico en la posición 7 por ácido glutámico :

Secuencia previa	M	V	L	S	P	A	D	K	T	N	P	D
Puntaje global	0	0	2	2	4	5	4	3	5	6	2	2
Puntaje total(previo)	35											
Secuencia mutada	M	V	L	S	P	A	E	K	T	N	P	D
Puntaje global	0	0	3	3	4	5	4	3	5	6	2	2
Puntaje total(posterior)	37											

Para este caso $Puntaje_{previo} = 35$ y $Puntaje_{posterior} = 37$. Usando un valor de $\beta = 1,5$ la ecuación 2.1 resulta en $P_{aceptar} = 0,26$. Para aplicar esta probabilidad dentro del método, obtenemos primero un valor *random* en el rango $[0,1]$, si el valor obtenido es menor que el calculado con la ecuación 2.1 entonces se acepta la mutación propuesta, caso contrario esta no es aceptada. En caso de NO aceptarse la mutación propuesta el algoritmo vuelve al paso 1, es decir, se vuelve a intentar una nueva mutación.

Dado que el objetivo final es obtener una secuencia que posea, de acuerdo con las herramientas bioinformáticas usadas en la evaluación, todas las características deseadas, el algoritmo finaliza cuando el puntaje resultante sea igual a cero, o cuando se cumple alguna de las condiciones de finalización especificadas en 4.2.5.

Capítulo 3

Análisis de las secuencias

En este capítulo se describen las evaluaciones que la herramienta permite realizar sobre las secuencias para detectar características deseadas y no deseadas durante el proceso iterativo de diseño. Se explica en general cuál es el objetivo de aplicar cada uno de los métodos y los fundamentos en los que se basan. Se detalla cómo se integran los distintos recursos en el contexto de nuestra aplicación, representando sus resultados en el esquema de puntajes propio de nuestro método.

El conjunto de evaluaciones que se describen corresponde a esta primera versión de nuestra herramienta. Por lo tanto no representa un conjunto exhaustivo ni definitivo, sino una primera etapa que deberá ser analizada y modificada de forma iterativa en función de los resultados obtenidos.

La definición del conjunto de evaluaciones utilizadas está directamente asociado a los fundamentos del método vistos en la sección 2.1. Por un lado, se tiene en cuenta la hipótesis general sobre el espacio de secuencias buscadas, que nos permite asumir que este espacio es relativamente grande con respecto al conjunto total de posibles soluciones. Por otro lado tenemos un conjunto de herramientas bioinformáticas que utilizan mecanismos y aproximaciones muy distintas y cuyos resultados pueden complementarse. Basándonos en estas propiedades, podemos pensar que la utilización de diferentes herramientas para la detección de las características de interés, aun cuando algunos resultados de estas puedan solaparse entre si, es una buena práctica al momento de definir el conjunto de herramientas utilizadas.

Por lo tanto, un aspecto que se repite es el solapamiento entre las características detectadas por las herramientas de evaluación, pensado con el fin de hacer mas exhaustiva la detección, además de políticas considerablemente abarcativas en cuanto a los métodos, con el fin de asegurar una mayor cobertura de las propiedades buscadas. Esta decisión permite dar mejores diseños finales, sin tener un aumento considerable en el tiempo de búsqueda. Como complemento de esta política, parte del trabajo

realizado durante el desarrollo de la herramienta consistió en implementar un código versátil que permita modificar fácilmente el conjunto de métodos de evaluación.

3.1. Propiedades conformacionales

Uno de los objetivos fundamentales de esta herramienta consiste en la restricción de elementos estructurales ordenados en la secuencia. Esto brinda al diseño resultante la flexibilidad requerida para la función de linker, adoptando una conformación intrínsecamente desordenada.

En primer lugar vamos a utilizar la herramienta IUPred, descrita en la sección 3.1.1, para intentar detectar aquellas posiciones que puedan generar interacciones favorables en el contexto de la secuencia, indicando una tendencia a adoptar estructuras plegadas, características de las proteínas globulares.

Por otro lado, utilizamos TMHMM (ver sección 3.1.2) para predecir la ocurrencia de segmentos transmembrana, los cuales presentan estructuras ordenadas con propiedades particulares que permiten identificarlos como módulos proteicos independientes.

Los agregados proteicos son otro tipo de módulos identificables que involucran la formación de una estructura tridimensional estable, en este caso mediante interacciones con otras unidades proteicas. Estas conformaciones agregadas pueden o no tener una estructura tridimensional ordenada pero, en todos los casos, limitan la flexibilidad requerida para una secuencia linker, generando, además, módulos proteicos insolubles que afectan a todo el sistema biológico. Debido a estas implicancias altamente negativas se evaluará detalladamente la tendencia a formar agregados dentro de la secuencia que estamos diseñando.

Existe actualmente una gran variedad de métodos existentes para predecir agregación amorfa y formación de fibras amiloides [18, 55, 89]. Cada método hace sus propias hipótesis e implementa predictores independientes, los cuales varían desde análisis muy simples (por ejemplo, análisis de la composición de aminoácidos) a métodos específicos más complejos, siendo la capacidad para formar hojas β una característica central de las evaluaciones, ya que es un denominador común de la formación de agregados.

En primer lugar, en nuestra herramienta, evaluaremos la tendencia a formar agregados utilizando TANGO (sección 3.1.3). Para evaluar específicamente la formación de fibrillas amiloides utilizaremos Waltz (descrito en la sección 3.1.5) y PASTA (sección 3.1.4). Por último, evaluamos la presencia de determinantes secuenciales que pueden indicar la formación de fibrillas amiloides, descrito en la sección 3.1.6.

3.1.1. IUPred: Análisis de tendencia al desorden

La herramienta IUPred [40] permite diferenciar aquellas secuencias que pueden formar suficientes interacciones favorables entre sus residuos, y por lo tanto podrían plegarse en una estructura globular, de aquellas no tienen esta capacidad y, por si solas, permanecen en una conformación desordenada. Esta diferencia nos permite, dentro de nuestro método, identificar las posiciones de la secuencia evaluada que tienen una mayor tendencia a formar conformaciones globulares, una propiedad que describimos como no deseada para secuencias linker.

El primer paso en el desarrollo de IUPred es definir un modelo simple para calcular la energía total de una estructura nativa de una proteína, a partir de los contactos que se encuentran en esta:

$$E = \sum_{ij=1}^{20} M_{ij} C_{ij} \quad (3.1)$$

donde M_{ij} es el potencial de interacción entre dos residuos de tipos i, j (sólo depende del tipo de residuos), y C_{ij} es el número de este par de residuos que se encuentran en contacto en la estructura.

Sin embargo, el método IUPred busca poder evaluar la contribución energética de cada aminoácido usando, únicamente, la composición de aminoácidos como parámetro. Esta evaluación tendrá la forma:

$$\frac{E_{estimada}}{L} = \sum_{ij=1}^{20} n_i P_{ij} n_j \quad (3.2)$$

donde n_i y n_j representan la frecuencia de residuos de tipo i y j , respectivamente, en la secuencia. P es la matriz de predicción de energía, que indica cómo la energía de un residuo de tipo i depende de la existencia de residuos de tipo j en el contexto.

De esta forma, sin conocer la estructura de la proteína, basamos el cálculo de la energía en valores estadísticos (matriz P) que pueden ser extraídos de una base de datos de proteínas globulares. Para calcular P , primero se desglosa la energía total de cada proteína en las contribuciones que hace cada tipo de aminoácido. Esto puede hacerse reusando el modelo de la ecuación 3.1, ya que conocemos las estructuras. La energía total por tipo de residuo se obtiene como:

$$e_i(\text{evaluada}) = \sum_{j=1}^{20} M_{ij} C_{ij} \quad (3.3)$$

Usando una aproximación similar a la de la ecuación 3.2, esta energía por tipo

de residuo se puede estimar mediante:

$$e_i(\text{estimada}) = N_i \sum_{j=1}^{20} P_{ij} n_j \quad (3.4)$$

donde N_i representa el número total de residuos de tipo i en la secuencia ($L * n_i$).

Cada valor P_{ij} de la matriz puede estimarse minimizando la diferencia entre los $e_i(\text{evaluada})$ y los $e_i(\text{estimada})$ para todas las estructuras de una base de datos de estructuras globulares. La función a minimizar es, entonces:

$$Z_i = \sum_k (e_i^k - N_i^k \sum_{j=1}^{20} P_{ij} n_j^k)^2 \quad (3.5)$$

donde k indica el índice de la estructura en la base de datos.

Una vez obtenidos los valores de la matriz P , tenemos un modelo completo que nos permite estimar la energía mínima de la estructura asociada a una secuencia, sin asumir ninguna conformación (3.2). Este modelo primero se puso a prueba sobre dos conjuntos de proteínas globulares e IDPs obteniéndose que la energía estimada para el conjunto de IDPs son menos favorables que las correspondientes al set de proteínas globulares, lo que está de acuerdo con la hipótesis que indica que las proteínas globulares tienen secuencias específicas con potencial para formar un gran número de interacciones favorables, mientras que las IDPs no. Utilizando esta separación significativa, lo que resta es transformar esta aproximación en un método para predecir el desorden a partir de la secuencia, para lo cual se transforma el valor de energía en un valor de probabilidad (*score* resultante, s_k).

La herramienta para el cálculo del score a partir de una secuencia está disponible a través de un servidor web [7, 39] o descargando la implementación y ejecutándola localmente [6]. En nuestro caso utilizaremos la segunda opción. Tal como se menciona en la información del servidor [39], los residuos que tengan un valor asociado de *score* mayor a 0.5 pueden ser tomados como desordenados. Por lo tanto, en nuestro método, los residuos que posean un *score* resultante menor a 0.5 tendrán un valor de 1 en el puntaje asociado a la posición. Por ejemplo, la evaluación de la secuencia VLKQTKGVGASGSFR con IUPred devuelve los valores de score que se ven en la figura 3.1

Utilizando nuestro esquema de evaluación, estos valores resultan en los siguientes puntajes:

Secuencia	V	L	K	Q	T	K	G	V	G	A	S	G	S	F	R
Evaluación con IUPred	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0

Disorder prediction score		
Position	Residue	Disorder Tendency
1	V	0.4476
2	L	0.4766
3	K	0.3667
4	Q	0.4256
5	T	0.4409
6	K	0.4409
7	G	0.4409
8	V	0.4409
9	G	0.4409
10	A	0.4409
11	S	0.4409
12	G	0.4831
13	S	0.5620
14	F	0.5419
15	R	0.5098

Figura 3.1

3.1.2. TMHMM: Secuencias transmembrana

Las proteínas de membrana son módulos proteícos con propiedades secuenciales distintivas y que, a pesar de su diferencia con las proteínas globulares, pueden adoptar una estructura tridimensional determinada, estabilizada por interacciones con el contexto hidrofóbico en el que normalmente se encuentran. Teniendo en cuenta este panorama más amplio, utilizamos el predictor TMHMM [63] para identificar, dentro de la secuencia que estamos evaluando, segmentos transmembrana ya que probablemente no hayan sido detectados por otros métodos de nuestra evaluación.

TMHMM es un método para detección de segmentos transmembrana que utiliza una aproximación mediante modelos ocultos de Markov (Hidden Markov Models, o HMM). Un HMM representa un modelo de Markov con estados no visibles. Es decir, el sistema es descrito por un modelo estocástico representado por estados y transiciones entre estos, las cuales están determinadas únicamente por el estado actual.

Describiendo el modelo de Markov que represente a la arquitectura de una proteína transmembrana se puede conocer si una secuencia desconocida pertenece a esta clase, evaluando si se adapta a este modelo. Para poder describir una proteína mediante un HMM se debe definir un conjunto de estados, cada uno correspondiente a una región o sitio específico de la proteína que se está modelando. Cada estado tendrá un valor de probabilidad asociado a cada uno de los 20 aminoácidos posibles. Además, se deben definir las transiciones posibles(y las probabilidades asociadas) entre los estados según la arquitectura que se está describiendo, por ejemplo si a partir de un estado pueden ocurrir nuevas instancias de este o si debe pasarse a un nuevo estado determinado. La distribución de probabilidades de los aminoácidos

en cada estado y la probabilidad de transición se derivan a partir de frecuencias observadas en un conjunto de proteínas conocidas que se adaptan al modelo

El modelo utilizado para describir la arquitectura de las proteínas en el método TMHMM (proteínas transmembrana) es cíclico y está compuesto por 7 estados distintos que representan: el núcleo de la hélice transmembrana, los dos extremos de ésta, el loop que se encuentra en el lado citoplasmático, dos loops en el lado no-citoplasmático, y un dominio globular en el medio de estos. En la figura 3.2 se ve la descripción gráfica de este. En [100] se describe como se procede al entrenamiento del HMM para poder obtener todos los parámetros asociados.

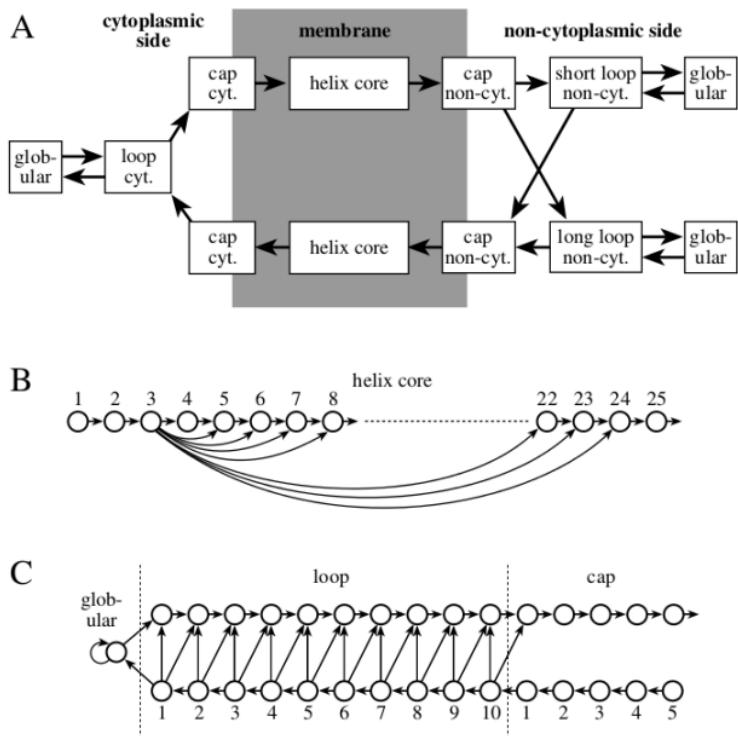


Figura 3.2: **A:** esquema general de la arquitectura asociada al modelo. En **B** y **C** se muestran en detalle algunos de los estados y cómo las transiciones entre estos permite definir los límites en las longitudes de las distintas regiones y sus propiedades. Figura extraída de [100]

Para evaluar secuencias se utiliza el algoritmo de Viterbi, el cual permite calcular cual es el camino más probable del modelo que adopta la secuencia a evaluar. Es decir, cual es la secuencia mas probable de estados del HMM (entre todas las posibles), que produce la secuencia de estados observados (aminoácidos de la secuencia). El resultado de la ejecución muestra cómo los residuos de la secuencia se ajustan a los diferentes estados del modelo de acuerdo a este recorrido más probable, clasificándolos según pertenezcan a regiones intracelulares, hélices transmembrana, o regiones extracelulares. Este método puede ser ejecutado como servicio web o des-

cargado como paquete de software en [15]. En nuestro caso realizamos la ejecución de manera local.

Nuestro interés está en detectar las regiones que forman los segmentos transmembrana, por lo tanto, el puntaje resultante será igual a 1 para aquellas posiciones que pertenezcan a estos segmentos y 0 para el resto. Por ejemplo, la ejecución de TMHMM a partir de la secuencia NFVLIGSFVAFFVITYFLE devuelve los siguientes resultados:

```
inside 1-1
TMhelix 2-18
outside 19-19
```

Por lo tanto, el puntaje resultante de la evaluación es:

Secuencia	N	F	V	L	I	G	S	F	V	A	F	F	V	I	T	Y	F	L	E
Puntaje TMHMM	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	

3.1.3. Tango

TANGO [47] es un método utilizado principalmente para predecir regiones de agregación β , desarrollado a partir de un modelo de mecánica estadística que define un espacio de fases incluyendo, además de estos agregados, conformaciones de β -turn, hélices- α y hebras- β . En nuestro sistema de evaluación, deseamos evitar tanto las conformaciones agregadas como cualquier otra estructura ordenada, por lo tanto utilizaremos TANGO para detectar las posiciones que tienen tendencia a encontrarse formando alguna de estas conformaciones.

De acuerdo con el modelo definido por TANGO, cada segmento de un péptido/proteína puede encontrarse en alguno de estos estados de acuerdo a una distribución de Boltzmann. Es decir, la frecuencia con la que el segmento se encuentra en alguno de los posibles estados, es relativa a la energía asociada a este, la cual se deriva de consideraciones empíricas y estadísticas. De esta forma, el método consiste de un algoritmo que simplemente calcula la función de partición y en base a eso predice las regiones con tendencia a agregación β .

Puntualmente, para el cálculo de la energía asociada a la conformación de hélice- α , se tienen en cuenta los parámetros definidos previamente en el desarrollo del método AGADIR [64]. Dentro de estos parámetros, principalmente se tiene en cuenta la interacción entre cadenas laterales de los pares de residuos $i, i+3$ e $i, i+4$.

Para el cálculo de la energía asociada a las conformaciones β , sólo se tiene en cuenta la energía de interacción entre las cadenas laterales de los residuos $i, i+1$ e $i, i+2$. Esta energía se deriva a partir de una base de datos de estructuras, calculando la relación entre frecuencias observadas y esperadas para cada par de aminoácidos

que se encuentra en este tipo de conformación. Asumiendo que la base de datos representa un sistema termodinámico en equilibrio, las frecuencias se relacionan con la energía de interacción de acuerdo a la siguiente ecuación:

$$\Delta G_{interacc-cadenas} = -RT \ln\left(\frac{f_{observada}}{f_{esperada}}\right) \quad (3.6)$$

Para el caso en que la conformación sea de β -turn, dado que los residuos no están fijos sino que pueden adoptar diferentes conformaciones, se agrega una penalización por el aumento de entropía. Por otro lado, en el caso de agregados β se asume que los residuos que adoptan esta conformación se encuentran completamente insertos en el núcleo y pagan el costo energético correspondiente. En este caso se calculan los diferentes parámetros energéticos, de hidrofobicidad, de solvatación, las interacciones electrostáticas y las interacciones por formación de puentes de hidrógeno.

A partir de las energías definidas para estas cuatro conformaciones, se calcula la función de partición resultante que permite definir las tendencias de cada residuo a poblar los distintos estados conformacionales evaluados. La salida que muestra TANGO se compone de un archivo con las siguientes columnas: Número de posición, residuo en esa posición, porcentaje en conformación hebra- β , porcentaje en conformación β -turn, porcentaje en conformación hélice- α , porcentaje de agregados β y porcentaje en agregación de hélices- α . Esta última columna no forma parte del modelo inicial que describimos y se calcula aparte, no representando un resultado confiable.

El método puede ser utilizado a través del servidor web en [14]. En nuestro caso solicitamos a los creadores especialmente una versión ejecutable para poder usarla de manera local. Una vez ejecutados y obtenidos los resultados para una secuencia, es necesario definir un valor de corte que determine la significancia de la tendencia obtenida. En [47] se definen dos intervalos de confianza para la predicción de regiones de agregación: segmentos con residuos que posean más de 5 % de tendencia agregación β (resultando en una alta certeza de predicción), y segmentos con residuos entre 0.2 % y 5 %. En nuestro caso utilizamos 1 % como valor de corte en esta primera implementación de la herramienta.

Dado que TANGO fue desarrollado con el objetivo de evaluar tendencias a formar agregados, no se evaluaron las predicciones asociadas a otros estados conformacionales. Para utilizar estas evaluaciones en nuestra implementación utilizaremos el mismo valor de corte que se obtuvo para la predicción de agregación β . Por lo tanto, el puntaje será igual a 1 cuando el residuo tenga una tendencia superior al valor de corte para la formación de agregados, hélices- α , hebras- β o β -turns, y un puntaje

igual a 0 en caso contrario. Por ejemplo, si evaluamos usando TANGO la secuencia AMAPVLYLQDKSS, obtenemos la siguiente tabla de valores:

Pos	Residuo	Beta	Turn	Hélice	Beta Aggregation	Helical Aggregation
01	A	0.2	0.0	0.000	0.000	0.000
02	M	0.2	0.0	0.000	0.000	0.000
03	A	0.2	0.0	0.000	0.000	0.000
04	P	0.0	0.0	0.000	0.544	0.544
05	V	1.6	0.0	0.000	10.620	10.620
06	L	2.3	0.0	0.000	10.620	10.620
07	Y	3.1	0.0	0.000	10.620	10.620
08	L	1.7	0.2	0.000	10.620	10.620
09	Q	1.0	0.2	0.000	10.203	10.203
10	D	0.3	1.8	0.000	0.000	0.000
11	K	0.3	1.8	0.000	0.000	0.000
12	S	0.2	1.6	0.000	0.000	0.000
13	S	0.2	1.6	0.000	0.000	0.000

El segmento en las posiciones 5-9 supera el porcentaje de agregación que utilizamos como valor de corte, mientras que el segmento 10-13 supera este valor en la tendencia a encontrarse en conformaciones de β -turn. Por lo tanto, el puntaje resultante de la evaluación es:

Secuencia	A	M	A	P	V	L	Y	L	Q	D	K	S	S
Evaluación con TANGO	0	0	0	0	1	1	1	1	1	1	1	1	1

3.1.4. PASTA

En nuestra herramienta, utilizamos PASTA [103] para identificar cuales son las regiones dentro de la secuencia que estamos evaluando, que podrían estabilizar una estructura de fibrillas amiloides. Para realizar esta predicción, el método realiza un cálculo de las energías de interacción entre los distintos segmentos, asumiendo que el mecanismo de las interaccion entre aminoácidos que lleva a la formación de láminas- β en proteínas globulares es el mismo que lleva a la formación de apilamientos de hebras- β en fibrillas amiloides.

La evaluación de la energía de interacción se realiza a partir de un potencial estadístico derivado de una base de datos de proteínas globulares. Para esto, se dividen las instancias encontradas de cada par $a-b$ de residuos en 4 categorías, según estén

interaccionando formando una hoja plegada- β en forma paralela (n_{ab}^p) o antiparalela (n_{ab}^a), o si no están participando de una estructura β y sus carbonos- α están a menos de 6,5 Å (n_{ab}^c contactos genéricos), o a más de 6,5 Å (n_{ab}^d , pares desordenados sin contacto). A partir de las frecuencias obtenidas se pueden derivar valores de energía asociados a las interacciones de a pares para los distintos estados, asumiendo que la base de datos analizada es un sistema en equilibrio termodinámico a temperatura constante para todas las proteínas. De esta forma, la probabilidad de cada par $a-b$ de encontrarse en un estado x se relaciona con el valor de la energía mediante el factor de Boltzman $p_{ab}(x) = e^{-E_{ab}^x}$.

Si podemos obtener una aproximación para $p_{ab}(x)$ se puede despejar el valor del potencial de interacción estadístico asociado (E_{ab}^x) para cada estado x , el cual corresponde a la diferencia en energía entre el estado x y el estado que se toma como referencia. Definiendo la probabilidad $p_{ab}(x)$ como la relación entre la frecuencia de interacciones observada para cada par y la esperada en el estado de referencia, y aproximando esta última como la frecuencia observada para todos los pares, se obtiene:

$$E_{ab}^x = -\log \left(\frac{\frac{n_{ab}^x}{n_{ab}}}{\frac{\sum_{ab} n_{ab}^x}{\sum_{ab} n_{ab}}} \right) \quad (3.7)$$

Para predecir los segmentos que pueden estar involucrados en la agregación de una proteína se prueban todos los emparejamientos posibles entre subsecuencias de ésta, en sentido paralelo y antiparalelo, y para cada uno se calcula la energía asociada, resultante de sumar todos los potenciales de interacción de a pares (E_{ab}) involucrados.

Sobre los valores de energía resultantes de estas sumas se aplica un punto de corte, siguiendo las recomendaciones analizadas en [113]. Puntualmente, utilizamos el esquema de valores que se describe como más específico, con un valor de $cut-off = -5$ y teniendo en cuenta sólo el emparejamiento de menor energía aun cuando existan otros con valores menores al $cut-off$. Sólo usamos el primero de los emparejamientos porque, al implementar un esquema iterativo, si los emparejamientos que se encuentran por debajo del valor de $cut-off$ persisten, se incrementan o se modifican, igualmente los tendremos en cuenta a todos en las próximas iteraciones.

El método PASTA es ejecutado de forma local mediante un script escrito en lenguaje Perl, obtenido directamente de sus desarrolladores, junto con las matrices

de energías. Si los resultados indican que algún segmento de la secuencia tiene una tendencia considerable a la formación de agregados, simplemente se asigna un puntaje de 1 a cada posición de este. La mayoría de las veces los emparejamientos de menor energía se corresponden con emparejamientos en forma paralela e *in-register* (PIRA), es decir, involucran el mismo segmento en dos moléculas que interactúan. En otros casos, el emparejamiento corresponde con dos regiones distintas de la misma molécula. En ese caso PATENA asigna el puntaje 1 a las posiciones de ambas regiones.

Para ejemplificar el proceso de evaluación, se muestra el resultado del evaluar la secuencia VTNVGGAVVTGVTAV. La ejecución de PASTA sobre esta secuencia devuelve: `pairing 0 PASTA energy -5.735704 length 13 between segments 1-13 and 1-13 parallel`

Este resultado indica que existe un emparejamiento de forma paralela, con un score de $-5,735704$, entre dos segmentos correspondientes a la subsecuencia VTNVGGAVVTGVT. El puntaje resultante, por lo tanto, es:

Secuencia	V	T	N	V	G	G	A	V	V	T	G	V	T	A	V
Evaluación con PASTA	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0

3.1.5. Waltz

Utilizamos Waltz [74] dentro de las evaluaciones de nuestra herramienta para predecir la tendencia de la secuencia a formar agregados amiloideos. Específicamente, el método Waltz determina la tendencia a formar agregados amiloideos en base a tres contribuciones: un componente derivado de información secuencial, un componente derivado de un conjunto de 19 propiedades fisicoquímicas, y un componente resultante de evaluar distintos residuos sobre un modelo estructural de la cadena carbonada de fibras amiloideas.

El primer componente se obtiene a partir del análisis secuencial de una bases de datos de hexapéptidos generadores de fibrillas amiloideas, los cuales han sido evaluados experimentalmente (AmylHex). Esta base de datos y el estudio realizado están basados en hexapéptidos porque la mayoría de las secuencias amiloideas disponibles hasta el momento tienen esta longitud. De esta forma, se ha tomado como una longitud representativa del núcleo que genera la formación de amiloideos, asumiendo que la inserción de estos segmentos es suficiente para inducir la conversión de todo el dominio de la proteína hacia una estructura agregada.

La base de datos utilizada, sin embargo, posee una alta redundancia secuencial ya que tiene una gran cantidad de entradas correspondientes a mutaciones puntuales del péptido STVIE. Para reducir esto, el conjunto inicial de la base de datos se utiliza

para obtener una PSSM inicial e identificar nuevas instancias de hexapéptidos, los cuales fueron evaluados experimentalmente para el desarrollo de la herramienta. A partir de este conjunto expandido de datos se realiza un alineamiento y se usa para generar una nueva PSSM, utilizando el método de puntuación por log-probabilidades (log-odd score). Este es el primer componente de la función de scoring de Waltz.

El segundo componente se deriva de analizar un conjunto de propiedades físicas sobre el conjunto de secuencias alineadas. Estas 19 propiedades se incluyen en la función como un score que se deriva de sumar, para cada posición y cada aminoácido, el producto entre la frecuencia de ese residuo y el valor normalizado de la cada propiedad. De esta forma, se obtiene un perfil de las propiedades físicas para cada posición.

El último componente de la función de scoring es una PSSM derivada del modelado estructural. Para obtener esta matriz se utiliza el campo de fuerzas FoldX [96], y la estructura tridimensional del péptido GNNQQNY en una fibrilla amiloide (conocido como formador de agregados amiloides, proveniente de Sup35). En primer lugar se mutan todas las posiciones a Alanina calculando el valor de ΔG para obtener finalmente el hexapéptido poli-A. Sobre este se comienzan a realizar todas las combinaciones posibles de mutaciones utilizando los 20 aminoácidos naturales, calculando ahora el valor de ΔG con respecto a este hexapéptido poli-A de referencia y el $\Delta\Delta G$ con respecto al péptido original. El valor del score específico para cada posición y cada aminoácido se obtiene de promediar los valores de energía fijando este aminoácido a la posición y combinando todos los demás aminoácidos posibles en el resto de las posiciones.

La función de scoring final resulta de la combinación lineal de los tres scores detallados:

$$S_{total} = \alpha S_{secuencial} + \beta S_{propiedadesFisicas} + \gamma S_{estructural} \quad (3.8)$$

El resultado de aplicar esta función puede obtenerse mediante un servidor web [16], el cual permite detectar todas las posiciones de una secuencia que forman parte de algún polipéptido y que supera cierto valor de *cut-off*. En el mismo servidor se definen dos opciones estándar para este valor de corte: alta especificidad, con un *cut-off* de 97, o alta sensibilidad, con un valor de *cut-off* de 79.

Para nuestro sistema de evaluaciones obtuvimos, de los desarrolladores del método, una versión de la PSSM correspondiente junto con un script en lenguaje Perl que

permite evaluar el score de cada posición. Ejecutando el script sobre la secuencia en cada evaluación, usando un valor de $cut-off = 79$ (alta sensibilidad), asignamos un puntaje igual a 1 en todas las posiciones que superen este valor. Por ejemplo, ejecutando Waltz sobre la secuencia VTNVGGAVVTGVT, se obtiene que el segmento 6-13 forma parte de hexapéptidos con score mayor al $cut-off$. Por lo tanto, el puntaje de la evaluación será:

Secuencia	V	T	N	V	G	G	A	V	V	T	G	V	T
Evaluación con Waltz	0	0	0	0	0	1	1	1	1	1	1	1	1

3.1.6. Determinantes secuenciales de fibras amiloides

En esta etapa de la evaluación utilizaremos un patrón secuencial extraído a partir de evaluaciones experimentales [35] con el fin de encontrar determinantes secuenciales para la formación de fibrillas amiloides sobre la secuencia linker que estamos evaluando.

El trabajo realizado para obtener el patrón secuencial que determina la formación de fibrillas amiloides se basa en un experimento de mutagénesis a partir de un péptido formador de fibrillas amiloides diseñado *de novo* en [34]. En este proceso se reemplazan sistemáticamente los residuos del péptido diseñado (STVIIIE) por todos los aminoácidos naturales, excepto Cisteína el cual no fue utilizado bajo la suposición que, por su similitud con Serina, las restricciones secuenciales para ambas serían similares.

El trabajo implica, luego, la evaluación experimental de los péptidos resultantes, para lo cual se monitorea la polimerización de hojas- β utilizando la técnica de diacroísmo circular (CD) y la detección de las fibrillas formadas mediante microscopía electrónica. Un análisis descriptivo se provee en [35], encontrándose una dependencia posicional con la formación de este tipo de estructuras de agregación, existiendo tanto posiciones muy tolerantes como restrictivas a las mutaciones. Dado que existen diferencias en los resultados experimentales según el estado de ionización de algunos residuos, se obtienen dos patrones distintos según el pH en el que se encuentran los péptidos. Los patrones resultantes son:

A pH ácido: $\{P\}_1 - \{PKRHW\}_2 - [VLS(C)WFNQE]_3 - [ILTYWFNE]_4 - [FIY]_5 - \{PKRH\}_6$
A pH neutro: $\{P\}_1 - \{PKRHW\}_2 - [VLS(C)WFNQ]_3 - [ILTYWFN]_4 - [FIY]_5 - \{PKRH\}_6$
Donde los {} indican residuos “prohibidos” en esa posición, y los [] representan aquellos que son “aceptados”, con respecto a la formación de amiloides. Los subíndices indican las posiciones correspondientes en el hexapéptido.

El patrón secuencial resultante no será capaz de detectar, por si sólo, todos los

motivos asociados con la generación de fibrillas amiloides, ya que es el resultado de un análisis sobre un espacio muy reducido de secuencias. En nuestro caso, sin embargo, es un elemento útil ya que es aplicado en conjunto con una serie de recursos adicionales dentro de un análisis secuencial exhaustivo.

Utilizaremos el patrón correspondiente a pH ácido por ser el más general de los dos. Para implementar la evaluación de la secuencia usando este patrón, se buscan instancias de la expresión regular asociada a éste dentro de la secuencia.. Para buscarlas se utiliza el módulo `re` de Python que permite, justamente, buscar instancias de una expresión regular sobre una secuencia. Todas las posiciones que pertenecen a instancias de este patrón tendrán un puntaje igual a 1. Por ejemplo, al evaluar la secuencia HPALFTIWHP se encuentra una instancia del patrón buscado en la subsecuencia ALFTIW, por lo tanto el puntaje correspondiente es:

Secuencia	H	P	A	L	F	T	I	W	H	P
Evaluación en busca de patrón secuencial	0	0	1	1	1	1	1	1	0	0

3.2. Elementos biológicamente funcionales

Para asegurar que el linker diseñado funcione únicamente como conector flexible entre dominios, debemos evaluar que sea biológicamente inerte. Esto implica que su secuencia no posea interacción alguna, evitando así cualquier interferencia con la expresión, utilización y actividad biológica de la proteína químérica. Realizamos la evaluación de posibles funcionalidades biológicas existentes en la secuencia a través de distintos métodos.

Utilizando la herramienta BLAST (descrita en 3.2.1) intentaremos detectar posibles regiones biológicamente funcionales infiriéndolas a partir de la similitud con proteínas naturales. En muchos casos, sin embargo, los determinantes secuenciales para la función biológica no están homogéneamente distribuidos a lo largo de la secuencia o son limitados en tamaño, por lo tanto no serán detectados por métodos de similitud secuencial. Para detectar este tipo de elementos funcionales utilizamos el recurso PROSITE (descrito en 3.2.2).

Más allá de estos, en la sección 1.2.3.3.3 describimos la existencia de motivos lineales cortos (short linear motifs, o SLiMs) como elementos funcionales con propiedades particulares. Para detectar SLiMs se utilizará principalmente el recurso ELM, cuya aplicación en nuestro método se describe en 3.2.3.

Además de los motivos lineales, en la sección 1.2.3.3.3 se describieron otros módulos funcionales que suelen estar contenidos en IDR/IDPs. Dentro de estos, los Mo-

REs son elementos que intervienen en procesos de señalización y reconocimiento entre proteínas y pueden distinguirse a partir de sus propiedades conformacionales características. Estos elementos son detectados en nuestra evaluación mediante la herramienta ANCHOR (descrita en 3.2.5)

Otro tipo de interacciones biológicas relevantes son aquellas mediadas por chaperonas. Dentro del complejo mecanismo de proteostasis celular, las chaperonas son elementos fundamentales para el correcto funcionamiento y calidad de las proteínas. El reconocimiento por parte de chaperonas implica la unión a ésta lo cual puede, además, interferir en la flexibilidad de la secuencia linker diseñada. Para intentar detectar en la secuencia de trabajo motivos asociados con el reconocimiento por parte de chaperonas utilizamos la herramienta Limbo, detallada en la sección 3.2.4.

3.2.1. BLAST

El método BLAST [19, 75] permite evaluar la similitud entre dos secuencias biológicas, tales como cadenas de aminoácidos correspondientes a proteínas o secuencias nucleotídicas. En nuestra herramienta utilizaremos BLAST para inferir la existencia de elementos funcionales en la secuencia linker que estamos evaluando a partir de una similitud considerable con secuencias naturales, las cuales asumimos que poseen una funcionalidad biológica.

Para realizar la comparación, BLAST requiere una secuencia de búsqueda (query) y una secuencia contra la cual comparar. La ventaja principal del método es que permite realizar, de forma muy eficiente, la comparación de una misma secuencia contra gran cantidad de secuencias contenidas en una base de datos. Para esto, BLAST utiliza un método heurístico no exhaustivo llamado word method o método de k-tuplas. Al ser un método heurístico, no está garantizado que encuentre los alineamientos óptimos con las secuencias de la base de datos, lo que sí ocurriría si se utilizara el algoritmo de Smith-Waterman [99]. Este último permite encontrar el alineamiento óptimo a expensas de un gran costo computacional.

En primer lugar, BLAST extrae de la secuencia query todas las subsecuencias de largo k (valor definido según el tipo de secuencias que se están comparando) y busca ocurrencias de éstas en la base de datos. A partir de estos alineamientos locales, se seleccionan aquellos con mayor valor de score y se extiende el alineamiento hacia la derecha e izquierda de las secuencias, hasta que ocurre una disminución en el score correspondiente. Por último, si existen extensiones de alineamiento que se encuentran en una misma secuencia estas pueden unirse. Finalmente, se muestra un alineamiento completo entre las secuencias con mayor score, junto con la significancia estadística del alineamiento, dependiente del largo de la secuencia query y el tamaño

de la base de datos.

La herramienta BLAST puede ser ejecutada de dos maneras. La opción más simple es a través del servidor web [17], lo cual implica ejecutar la búsqueda remotamente con retardos importantes para obtener los resultados. La segunda opción es realizar la búsqueda de forma local, para lo cual es necesario tener disponible el paquete de software provisto por NCBI [13], junto con la base de datos sobre la cual se quiere hacer la comparación. Esta es la opción que utilizamos en nuestra herramienta y, en nuestro caso, realizamos la búsqueda sobre la base de datos UniProtKB [23]. Utilizamos 0.01 como valor de *cutoff*, es decir, todas las secuencias encontradas que tengan un e-value menor a 0.01 serán consideradas como significativamente similares. Todas las posiciones que, en el alineamiento final, sean idénticas en ambas secuencias, tendrán un puntaje igual a 1 en nuestro sistema de evaluación.

Como ejemplo de aplicación realizamos la búsqueda de la secuencia MVLSPADKTNVKGGWGKV, encontrando la secuencia MVLSPADKTNVKAAGKGV con un e-value de 7e-09. El alineamiento entre estas dos secuencias y el puntaje resultante asignado por nuestro método es:

Secuencia	M	V	L	S	P	A	D	K	T	N	V	K	G	G	W	G	K	V
Alineamiento Hit	M	V	L	S	P	A	D	K	T	N	V	K	-	-	W	G	K	V
Puntaje BLAST	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1

La búsqueda BLAST, aun siendo un método heurístico y altamente optimizado, tiene un tiempo de ejecución considerablemente alto con respecto al resto de las herramientas descritas. Por otro lado, para encontrar un *hit* estadísticamente significativo la secuencia debe tener, generalmente, un largo considerable. Esto se debe a que el valor de *e-value* depende, en parte, del largo de la secuencia *query*.

Gran parte de las evaluaciones realizadas en nuestro método no cumplen esta característica de longitud. De esta forma, para reducir el tiempo de ejecución global de nuestro método, la evaluación usando BLAST se realiza de forma separada. Para esto, en primer lugar se realiza el proceso iterativo de mutaciones utilizando todo el conjunto de herramientas de evaluación excepto BLAST. Una vez que se alcanza una secuencia con score = 0 según ese esquema de evaluación, se realiza el mismo proceso utilizando exclusivamente BLAST para las evaluaciones. La iteración global termina cuando, usando el total de las herramientas aquí descritas, la secuencia alcanzada tiene un score = 0.

3.2.2. Prosite

PROSITE [9, 97] es un recurso que agrupa una gran cantidad de motivos secuenciales de diversas características, permitiendo anotar e identificar regiones conservadas en secuencias de proteínas. En nuestra implementación utilizamos este recurso para extender la búsqueda de elementos funcionales en la secuencia.

De forma resumida, se puede definir a PROSITE como una colección anotada de motivos biológicamente significativos, dedicada a la identificación de familias y dominios de proteínas. Esta base de datos contiene información derivada de alineamientos de múltiples secuencias homólogas. Los motivos resultantes se describen usando dos métodos distintos, cada uno con sus ventajas y desventajas.

La primera forma de describir los motivos es a través de patrones secuenciales (utilizando expresiones regulares como se mostró para ELM, sección 3.2.3), en los cuales se tiene en cuenta solo la información de los residuos más significativos, descartando el resto. La búsqueda de un patrón en una secuencia da un resultado cualitativo: hay una coincidencia o no la hay. Si hay una sustitución en alguna de las posiciones de la secuencia el patrón no coincide, independientemente del tipo de sustitución que ocurrió.

Otra forma para describir los motivos es mediante perfiles (o matrices de pesos). Estos pesos proveen valores numéricos para cada posible coincidencia o sustitución cuando se busca el motivo en una secuencia. De esta forma, al utilizarlos en la búsqueda de un motivo, funcionan como descriptores cualitativos que consideran la similitud global en toda la longitud secuencial de un dominio o proteína. Un motivo puede ser encontrado en una secuencia que posee una sustitución en una posición conservada si el resto de la secuencia tiene un nivel de similitud suficientemente alto. Estas propiedades dan una mayor sensibilidad a los perfiles con respecto a los patrones, permitiendo encontrar dominios o familias con alta divergencia que solo tienen unas pocas posiciones muy conservadas.

Diversas búsquedas relacionadas con patrones anotados en PROSITE se pueden hacer a través de la herramienta ScanProsite [12, 33], la cual permite escanear secuencias para buscar ocurrencias de los motivos, buscar motivos en una base de datos entera de secuencias, o buscar motivos propios del usuario en una secuencia. Esta herramienta se encuentra disponible para descargar, junto con la base de datos completa de motivos secuenciales, lo que permite realizar la búsqueda de forma local.

Para encontrar motivos secuenciales en la secuencia que evaluamos, es posible escanearla utilizando patrones y/o perfiles, y variar también los límites usados en la detección de los perfiles. En nuestro caso utilizaremos patrones para realizar la

búsqueda, principalmente porque el tiempo de búsqueda utilizando perfiles es de aproximadamente 10 veces el que demora la búsqueda mediante patrones y, si bien esta diferencia es despreciable cuando se hacen búsquedas sobre una sola secuencia, al realizar evaluaciones en una gran cantidad de iteraciones la diferencia se vuelve significativa. En segundo lugar, consideramos que la sensibilidad provista por los patrones es suficiente para los fines buscados en nuestra herramienta, al menos inicialmente. De todas formas, es uno de los aspectos a evaluar con mayor profundidad a futuro, por lo que no se descarta extender la búsqueda para poder utilizar perfiles, al menos de forma opcional para el usuario.

Un aspecto relevante de la base de datos PROSITE es que, si bien se ha orientado hacia la anotación de dominios globulares por sobre motivos lineales, existen anotaciones que representan motivos lineales cortos, lo cual trae dos consecuencias. En primer lugar, para cualquier búsqueda, pueden ocurrir una gran cantidad de falsos positivos debido a las propiedades intrínsecas de estos. En segundo lugar, debido al contexto en el cual la estamos aplicando, donde también se buscan SLiMs mediante el recurso ELM, puede haber un solapamiento con los resultados obtenidos entre ambas herramientas.

Usando la secuencia VKTCLALGVDINTCD ejemplificaremos el proceso de evaluación. Mediante la ejecución de ScanProsite se identifica que esta secuencia contiene el patrón PS00008 (MYRISTYL N-myristoylation site <http://prosite.expasy.org/cgi-bin/prosite/nicedoc.pl?PS00008>) ubicado en la subsecuencia GVDINT (posiciones 8-13), por lo tanto el puntaje correspondiente es:

Secuencia	V	K	T	C	L	A	L	G	V	D	I	N	T	C	D
Evaluación global ELM	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0

3.2.3. ELM

El recurso de motivos lineales eucariotas (ELM) [38, 86] fue establecido con la misión de recolectar, anotar y clasificar motivos lineales cortos (conocidos como LMs, ELMs, SLiMs o MiniMotifs y detallados en la sección 1.2.3.3.3). Lo utilizaremos para detectar instancias de estos elementos dentro de la secuencia linker que estamos evaluando, algo que, debido a las propiedades intrínsecas de los SLiMs, no sería posible sin un recurso con datos curados manualmente de la literatura y que están completamente a disposición de la comunidad científica, como es ELM.

Este recurso provee actualmente una completa base de datos con motivos organizada jerárquicamente: en el nivel superior se tiene un conjunto de tipos (actualmente hay un total de 6 tipos diferentes). Cada tipo agrupa un conjunto de clases, y ca-

da clase define la especificación de un dominio o familia de dominios de péptidos, los cuales se describen mediante una expresión regular (cadena de caracteres para describir un patrón secuencial [4]) representativa de la secuencia que los compone. Cada clase contiene al menos una instancia anotada, donde cada instancia representa una secuencia determinada experimentalmente que se ajusta a la expresión definida para la clase. El énfasis está puesto en la validación experimental que ha sido realizada sobre estas secuencias, logrando un proceso de curación manual a partir de la literatura con las instancias que son ingresadas en la base de datos.

Dado que los motivos suelen tener solo un pequeño número de posiciones fijas, es normal que las búsquedas resulten en una gran cantidad de falsos positivos. Es por esto que el recurso también provee opciones para filtrar los resultados según la especie, el compartimento en el cual se va a encontrar la secuencia, etc. De todas formas, la validación final de la funcionalidad del motivo debe ser siempre realizada experimentalmente.

Este recurso puede ser utilizado directamente a través del servidor web [3], el cual provee una herramienta para encontrar, en una secuencia ingresada por el usuario, instancias de los motivos contenidos en la base de datos. Otra forma de realizar esto es haciendo una búsqueda local. Para ésto es necesario descargar la base de datos de expresiones regulares correspondientes a los motivos y, a partir de estos datos, realizar la búsqueda de cada expresión regular sobre la secuencia con la que estamos trabajando. Esta última forma de detección es la que utilizaremos para nuestra evaluación, mediante el módulo `re` de Python que permite buscar instancias de una expresión regular sobre una secuencia. La base de datos utilizada corresponde a la versión con fecha 1-3-2015 obtenida de [3].

La búsqueda de motivos sobre la secuencia resulta en un conjunto de subsecuencias correspondientes a cada instancia encontrada las cuales pueden estar solapadas, es decir, cada posición de la secuencia puede estar contenida en 0,1 o más instancias encontradas. Cada una de estas subsecuencias se trata de forma individual y, por cada posición de ésta, se suma 1 al puntaje asociado. A modo de ejemplo, si estamos trabajando con la secuencia PSKPLRGNAMVGL, el resultado de la búsqueda indica 2 motivos encontrados:

Clase ELM	Expresión regular	Instancia (ubicación)
LIG_SH3_2	P..P.[KR]	PSKPLR (1-6)
DOC_MAPK_1 MAPK	[KR]{0,2}[KR].[0,2}{KR}.{2,4}[ILVM].[ILVF]	KPLRGNAMVGL(3-13)

Cada instancia encontrada implica una posible funcionalidad mediada por un motivo lineal, por lo tanto, se aumenta en 1 el valor del puntaje a las posiciones involucradas para cada instancia por separado. El puntaje resultante de este paso

es:

Secuencia	P	S	K	P	L	R	G	N	A	M	V	G	L
LIG_SH3_2 (1-6)	1	1	1	1	1	1	0	0	0	0	0	0	0
DOC_MAPK_1 MAPK (3-13)	0	0	1	1	1	1	1	1	1	1	1	1	1
Evaluación global ELM	1	1	2	2	2	2	1	1	1	1	1	1	1

3.2.4. Limbo: Interacción con chaperonas

El método Limbo [107] es un predictor de sitios de unión a la chaperona DnaK, representante de la extensa familia Hsp70, que se especializa en la unión de regiones hidrofóbicas expuestas, generalmente presentes en polipéptidos desplegados. La unión de chaperonas a una secuencia gobierna una gran variedad de procesos tales como translocación, replegamiento y degradación de la proteína reconocida, además de la activación de un amplio rango de proteínas asociadas. El objetivo de aplicar esta herramienta en nuestro esquema de evaluación es detectar la presencia sitios de unión a chaperonas en la secuencia linker, evitando el reconocimiento y la interacción con proteínas en el sistema biológico de expresión o aplicación.

El método Limbo se desarrolló utilizando una combinación de información secuencial y estructural para analizar el perfil de secuencias que se unen a la DnaK. El primer paso en el desarrollo del método es crear un predictor basado exclusivamente en información secuencial. Para armar el set de aprendizaje usado, se realizaron una serie de ensayos experimentales en los cuales un conjunto de péptidos se inmoviliza en membranas de celulosa y luego de la incubación con DnaK se detectó la unión a estos mediante un anticuerpo específico.

Los péptidos usados en los ensayos de unión se obtienen de 3 conjuntos distintos. El set inicial se obtiene de predicciones basadas en el uso de TANGO [47] (ver sección 3.1.3), asumiendo que la DnaK se une a secuencias propensas a formar agregados. Usando los resultados obtenidos, se desarrolla un predictor simple de unión a DnaK, el cual es utilizado para la búsqueda de sitios de unión a chaperonas en el proteoma de *E.coli*, las cuales conforman el segundo set. El último conjunto corresponde a dos péptidos supuestamente reconocidos por la DnaK, ubicados en el factor σ de la RNA-polimerasa. Estos péptidos, luego de ser sometidos al ensayo de binding, se dividen en dos conjuntos según los resultados obtenidos: los que son reconocidos por la DnaK y lo que no.

El set de aprendizaje final se construye obteniendo primero todas las subsecuencias de 7 residuos posibles. Del set de péptidos de no-unión todos los posibles heptapéptidos formados son seleccionados, para dar el set de resultados negativos. Por su parte, del set de péptidos de unión sólo se incluyen en el set de péptidos po-

sitivos a aquellos que dan una mayor energía de unión a DnaK, evaluada mediante el campo de fuerzas de FoldX. A partir de los conjuntos de heptapéptidos positivos y negativos se obtienen dos matrices PSSM y una matriz final se obtiene restando los valores de score de la PSSM de no-unión a los valores encontrados en la PSSM positiva.

De manera independiente se contruyó un perfil secuencial a partir de un análisis puramente estructural de unión a DnaK. Para esto se hizo primero un análisis de mutaciones posicionales sobre el heptapéptido unido a la DnaK utilizando la estructura de ésta (cristalizada junto con un heparpéptido) y el campo de fuerzas FoldX. En primer lugar se mutaron todas las posiciones a Alanina y luego cada posición fue mutada individualmente a cada uno de los restantes 19 aminoácidos. Se obtuvieron así los valores de $\Delta\Delta G$ para cada residuo y cada posición, donde los valores más negativos indican una mejor unión a la DnaK. Por lo tanto, la PSSM se obtiene del negativo correspondiente a cada valor de $\Delta\Delta G$.

Los valores de todas las matrices son optimizados mediante un algoritmo de validación cruzada que permite eliminar algunos péptidos de los conjuntos de aprendizaje. Los perfiles representados en las PSSMs obtenidas del análisis secuencial y del análisis estructural por separado se combinan para dar el predictor final.

La versión final de la matriz y un script implementado en Python para realizar el análisis fue provisto por los desarrolladores de Limbo. El valor de corte utilizado es 11.08 (valor por defecto). Para utilizar el método de manera local se corre el script provisto y se pasa como parámetro el archivo contenido las secuencias a analizar en formato fasta. El resultado contiene una lista de heptapéptidos cuyo score calculado supera el valor de corte. Por ejemplo, analizando la secuencia DLWKLLPENNVLSP, el resultado obtenido es:

```
1 DLWKLLP 11.7190683353
3 WKLLPEN 23.0607028837
```

Este resultado indica que los heptapéptidos DLWKLLP y el WKLLPEN, poseen valores de score superior al valor de corte (11.7190683353 y 23.0607028837, respectivamente). Utilizando nuestro sistema de evaluación, se aumenta en 1 el puntaje de todas las posiciones asociadas a cada (debemos tener en cuenta que algunas posiciones pueden sumar más de 1 ya que los heptapéptidos pueden estar solapados). A partir de los resultados anteriores, el puntaje de la evaluación es:

Secuencia	D	L	W	K	L	L	P	E	N	N	V	L	S	P
Evaluación con Limbo	1	1	2	2	2	2	2	1	1	0	0	0	0	0

3.2.5. ANCHOR: predicción de MoREs

La herramienta ANCHOR[77] busca identificar una clase especial de segmentos desordenados que son capaces de experimentar un proceso de *binding & folding*, propio de la unión a una proteína globular. El objetivo de aplicar esta herramienta como parte de nuestra evaluación se debe, principalmente, a que estos procesos de reconocimiento molecular suelen estar asociados a la señalización de distintos procesos biológicos, algo que destacamos como negativo para un linker ya que podría afectar el normal funcionamiento del sistema biológico en el que se encuentra. De esta forma, intentamos detectar cuales son las posiciones que componen estos elementos en la secuencia que estamos evaluando.

Para identificar estos elementos, el método reutiliza el modelo definido para implementar la herramienta IUPred (ver sección 3.1.1), en la cual se logra estimar la energía de interacción asociada a cada posición basándose en el tipo de aminoácido y la composición del contexto más próximo. Usando este modelo se pueden calcular, no solo la capacidad de interacción intramolecular, sino también la energía de interacción en el contexto de la unión a una proteína globular. Mediante la diferencia entre estos valores, se pueden identificar cuales son los segmentos que experimentan interacciones favorables en este nuevo contexto y que podrían estabilizar el plegamiento y unión con respecto a la conformación desordenada.

Los segmentos buscados deben poseer residuos con propiedades específicas:

1. El residuo debe pertenecer a una región larga desordenada, es decir, fuera de cualquier dominio globular.
2. En el estado aislado, el residuo no debe ser capaz de formar uniones favorables con sus vecinos cercanos que le permitan plegarse.
3. El residuo tiene una ganancia neta de energía proveniente de la interacción con proteínas globulares.

Cada uno de estos criterios está asociado a un valor numérico. La predicción de los segmentos buscados se basa en una combinación derivada directamente de estos. La ecuación asociada tendrá entonces tres componentes que se combinan linealmente:

1. El primer componente resulta de promediar los valores de *score* obtenidos directamente de IUPred, en una ventana de tamaño w_1 (que deberá definirse como parte de los ajustes de este predictor) alrededor de cada residuo. Esto evalúa la tendencia al desorden que tiene el entorno de cada residuo, separando regiones desordenadas de posiciones puntuales que puedan tener cierta

tendencia al desorden.

$$S_k = \frac{1}{N} \sum_{j=b_{lower}}^{b_{upper}} score_j \quad (3.9)$$

donde N es el número de aminoácidos efectivamente contenidos en la ventana para obtener el promedio del residuo con índice k , y b_{lower} - b_{upper} los límites de esta. $score_j$ es el valor obtenido directamente de IUPred para el residuo en la posición j .

2. El segundo componente evalúa la ganancia de energía que tendrá el residuo al formar interacciones de a pares con los vecinos contenidos dentro de una ventana de tamaño w_2 . La ecuación asociada es idéntica a la obtenida para IUPred (ecuación 3.2), pero el valor de la ventana se redefine como parámetro, el valor del cual será ajustado al nuevo predictor de segmentos ANCHOR.
3. El tercer componente evalúa la ganancia de energía que tendrá el residuo al formar interacciones de a pares con una proteína globular, con respecto a la formación de contactos únicamente entre los vecinos (componente 2). Para hacer esta evaluación, se reutiliza el modelo de IUPred, pero ahora, la composición del contexto con el cual se dan las interacciones estará dado por la composición de una proteína globular hipotética. Para esto se utiliza la frecuencia de aminoácidos estándar en estas proteínas. La diferencia resultante entre la interacción con los vecinos propios de la secuencia y este nuevo valor calculado será:

$$E_i^{ganancia,k} = E_i^{intra,k} - E_i^{globular,k} \quad (3.10)$$

donde $E_i^{intra,k}$ y $E_i^{globular,k}$ representan la energía de interacción de a pares asociada a cada posición, y se calculan usando nuevamente la ecuación 3.2, con las frecuencias de aminoácidos del contexto de la posición k (en el primer caso) y las frecuencias estándar de proteínas globulares (en el segundo caso).

La ecuación para cada posición k de la secuencia, resultante de la combinación de criterios es:

$$I_k = p_1 S_k + p_2 E_i^{intra,k} + E_i^{ganancia,k} \quad (3.11)$$

Varios de los parámetros de este nuevo modelo ya fueron determinados previamente utilizando datos conocidos de estructuras de proteínas globulares (durante el desarrollo de IUPred). Queda determinar cuál es el peso de cada uno en el valor total (coeficientes de la combinación lineal) y los tamaños de las ventanas (w_1 y w_2 , correspondientes a los componentes 1 y 2).

Para determinar los valores óptimos de estos parámetros se utilizaron dos conjuntos de datos: un conjunto negativo compuesto por cadenas de proteínas globulares y un conjunto de resultados positivos compuesto por complejos formados por segmentos desordenados unidos a proteínas globulares.. Cabe destacar que no es posible entrenar el predictor utilizando un conjunto de proteínas desordenadas que se sepa que no forman uniones con proteínas globulares, principalmente porque no existe método preciso para comprobar que esto no ocurre. A pesar de esta condición, y que la cantidad de datos del conjunto de resultados positivos es considerablemente limitada, una ventaja del método es que contiene sólo 5 parámetros para los cuales se deben evaluar sus valores óptimos.

Como se define en [41], las regiones que tienen un $score > 0,5$ se puede tomar como potenciales segmentos de unión y, por lo tanto, asignaremos a cada residuos en estas un puntaje asociado = 1. Por ejemplo, la evaluación de la secuencia TFSLWKPENMLSPDD con ANCHOR devuelve los valores de $score$ mostrados en la figura 3.3.

Position Specific Score		
Position	Residue	ANCHOR Probability
1	T	0.8298
2	F	0.8048
3	S	0.7816
4	L	0.7601
5	W	0.6860
6	K	0.5037
7	P	0.3722
8	E	0.2968
9	N	0.2518
10	M	0.2283
11	L	0.1801
12	S	0.1501
13	P	0.1159
14	D	0.1108
15	D	0.1162

Figura 3.3

Las posiciones 1-6 tienen asociados valores $> 0,5$. Aplicando nuestro esquema de evaluación, estos resultados se traducen en la siguiente tabla de puntajes:

Secuencia	T	F	S	L	W	K	P	E	N	M	L	S	P	D	D
Evaluación con ANCHOR	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0

3.3. Otros propiedades evaluadas

Como parte de la herramienta se provee la opción de realizar evaluaciones adicionales sobre la secuencia que permiten restringir la búsqueda más allá de las propiedades intrínsecas que debe poseer una secuencia linker. Las distintas características

“extra” que se permiten definir sobre la secuencia complementan el objetivo principal de la herramienta que es proveer secuencias que puedan ser utilizadas en el contexto de desarrollos experimentales de biología molecular.

Al no formar parte del conjunto de propiedades evaluadas de forma estándar, se debe indicar la aplicación de estas opciones mediante parámetros al iniciar la ejecución, tal como se indica en la sección [4.2.4](#) del manual.

3.3.1. Carga neta de la secuencia

Esta opción permite que el usuario defina una carga neta deseada para la secuencia linker resultante lo que implica que, en cada paso, se evalúe la secuencia con respecto a la carga. Para esta evaluación se tienen en cuenta tres categorías de aminoácidos: con carga positiva (K,R,H), carga negativa (E,D) y neutros. Al evaluar una secuencia con respecto a la carga neta, puede ocurrir alguno de los siguientes casos:

- Si la carga neta deseada y la actual son iguales, todos los puntajes serán 0.
- Si la carga neta a la que se está apuntando es más positiva que la actual, los puntajes resultantes son:
 - aminoácidos con carga negativa = 2
 - aminoácidos con carga positiva = 0
 - aminoácidos neutros = 1

Por ejemplo, si se está evaluando la secuencia VKTCLALGVDI (carga neta=0), y el usuario indicó como requerimiento una carga neta objetivo igual a +2,

Secuencia	V	K	T	C	L	A	L	G	V	D	I
Evaluación carga neta	1	0	1	1	1	1	1	1	1	2	1

- Si la carga neta a la que se apunta es más negativa que la actual, los puntajes resultantes son:
 - aminoácidos con carga negativa = 0
 - aminoácidos con carga positiva = 2
 - aminoácidos neutros = 1

Por ejemplo, si se está evaluando la secuencia VKTCLALGVDI (carga neta=0), y el usuario indicó como requerimiento una carga neta objetivo igual a -2:

Secuencia	V	K	T	C	L	A	L	G	V	D	I
Evaluación carga neta	1	2	1	1	1	1	1	1	1	0	1

3.3.2. Absorción UV

El objetivo de esta opción es que la secuencia resultante no posea ningún aminoácido que absorba en el rango del UV (W,Y,F). En este caso, ya que se espera que el resultado final no posea ninguno de estos residuos, directamente se eliminan de la composición de aminoácidos que se utiliza para reemplazar las posiciones mutadas. Para eliminar los residuos que se encuentran en la secuencia, se asigna un puntaje de 1 a aquellas posiciones que contengan residuos absorbentes en el rango UV. Por ejemplo, si estamos evaluando la secuencia VYTCLALGWDI:

Secuencia	V	Y	T	C	L	A	L	G	W	D	I
Evaluación UV	0	1	0	0	0	0	0	0	1	0	0

Capítulo 4

Manual de uso

El método descrito en los capítulos previos se encuentra disponible bajo el nombre de PATENA. El origen del nombre se encuentra en la descripción del método, metafóricamente, como un mecanismo para “limpiar” secuencias, eliminando elementos funcionales y estructurales. De esta forma, las secuencias resultantes podrían describirse usando la expresión popular “limpias como una patena”. Esta expresión hace alusión al platillo de metal en el que se pone la hostia durante la celebración eucarística (<https://es.wikipedia.org/wiki/Patena>), el cual se mantiene siempre limpio y brillante.

4.1. Instalación

El software fue desarrollado para correr en un entorno Linux. Para poder instalar PATENA se necesita primero tener instalados los siguientes paquetes de software:

- Python >2.7 [10]
- BioPython [2]
- Perl [8]
- Blast software command line [13] (asegurarse que el ejecutable esté en el path del usuario)
- Git [5] (necesario sólo para la instalación)

Una vez que se cumplen todos los requerimientos, el proceso de instalación de la herramienta requiere dos pasos:

1. `git clone https://github.com/ieguinoa/patena`
2. `source install.sh`

Una vez instalado, la ejecución se realiza mediante:

`python patena.py [parámetros de ejecución]`

4.2. Parámetros de ejecución

4.2.1. Secuencia inicial

4.2.1.1. Secuencia inicial random

Por default, la ejecución comenzará a partir de una secuencia random generada utilizando la composición designada para la ejecución (ver 4.2.2). La longitud predefinida para la secuencia es de 10 aminoácidos. El usuario puede modificar este valor mediante el parámetro `length` de la forma:

`--length [longitud deseada]`

Por ejemplo:

`--length 30` *Inicia la ejecución con una secuencia random de largo = 30*

4.2.1.2. Secuencia inicial definida por el usuario

Para definir una secuencia inicial específica se debe usar el parámetro `seq` de la forma:

`--seq [secuencia]`

Por ejemplo:

`--seq AAHHWWLLLLHHGGG` *Inicia la ejecución con la secuencia AAHHWWLLLLHHGGG*

En caso de especificarse una secuencia inicial no es necesario definir longitud y, si se define, es ignorada.

4.2.1.3. Secuencias flanqueantes

Es posible especificar segmentos en uno o ambos extremos que, si bien no serán parte del linker ni serán mutados, si se tendrán en cuenta a la hora evaluar la secuencia. Para indicar estas secuencias flanqueantes se proveen los parámetros `left` y `right` los cuales definen secuencias hacia los extremos N- y C-terminal respectivamente. El formato para especificarlas es:

`--left [secuencia flanqueante N terminal]`

`--right [secuencia flanqueante C terminal]`

Por ejemplo:

--left AHWLLHHGG *Se asume que a la izquierda del linker se encuentra la secuencia AHWLLHHGG*

4.2.2. Composición de la secuencia

4.2.2.1. Composición estándar

Si no se define ningún parámetro al respecto, la composición *default* que se usa corresponde a la composición estándar obtenida de SwissProt [1].

En caso que se haya seleccionado la opción de secuencia silente en UV (ver sección 4.2.4.2) la composición es igual a la estándar pero los residuos que absorben en el rango del UV (W,Y,F) tendrán una frecuencia igual a 0.

4.2.2.2. Composición definida por el usuario

Es posible que el usuario defina una frecuencia específica para uno o más aminoácidos. La frecuencia se debe especificar en forma de porcentaje y cada aminoácido tiene un parámetro asociado para definir su frecuencia. Esto parámetro se corresponde con la nomenclatura de una sola letra asociada al aminoácido. Ejemplos de esto son:

--a [frecuencia Alanina]
--m [frecuencia Metionina]
--s [frecuencia Serina]

En caso que no se definan las frecuencias para todos los 20 tipos de aminoácidos, el resto tendrá la frecuencia estándar. El único requerimiento es que la suma total de las frecuencias sea menor o igual a 100. A modo de ejemplo:

--a 50 --l 10 --g 13.3

Esta definición de la composición generará una secuencia inicial con 50 % de Alanina, 10 % de Leucina y 13.3 % de Glicina. El porcentaje restante tendrá la distribución correspondiente a la composición estándar.

En caso que también se haya seleccionado la opción de secuencia silente en UV (ver sección 4.2.4.2), los residuos que absorben en el rango del UV (W,Y,F) tendrán una frecuencia igual a 0, independientemente del valor que se especificó.

4.2.3. Definición del parámetro Beta

Basándose en las evaluaciones realizadas (ver sección 5.2), el valor estándar de Beta está definido como 1.0. Es posible modificar este valor para ejecuciones individuales mediante el parámetro **beta** de la forma:

--beta [valor deseado de Beta]

Por ejemplo:

--beta 2.3 *Inicia la ejecución con un valor de beta = 2.3*

4.2.4. Evaluación de la secuencia

Si no se especifica ningún parámetro adicional, el set de evaluaciones estará compuesto de aquellas herramientas detalladas las secciones 3.1 y 3.2. El usuario puede quitar, para una ejecución individual, alguna de las herramientas del conjunto de evaluación. Para esto, cada herramienta tiene asociado un parámetro que permite quitarla del esquema de evaluación. Los parámetros disponibles son: --noblast, --notango, --noelm, --noiupred, --noanchor, --noprosite, --nolimbo, --notmhmm, --nopasta, --nowaltz, --noamyloidpattern

Por ejemplo:

--noblast --nowaltz

La evaluación se realiza usando todas las herramientas excepto BLAST y Waltz

El usuario puede, además, agregar evaluaciones correspondientes a la carga neta y a la absorción UV de la secuencia, las cuales se detallan a continuación.

4.2.4.1. Evaluación de carga neta

La carga neta deseada para la secuencia resultante puede ser definida mediante la opción `netcharge` de la forma:

--netcharge [carga neta deseada]

Por ejemplo:

--netcharge +3 *La secuencia resultante tendrá carga neta = +3*

El valor absoluto de la carga buscada debe ser menor o igual que la longitud de la secuencia, de otra forma no podría alcanzarse nunca la carga neta objetivo. Esta condición se evalúa al inicio de la ejecución y se informa si no es posible continuar.

4.2.4.2. Evaluación de silente en UV

Es posible restringir el conjunto de secuencias resultantes a aquellas que no absorban en el rango de UV, es decir, que no posean residuos Y, W o F. La forma de indicar esto es mediante la opción `--uvsilent`(sin parámetros)

4.2.5. Condición de finalización

Por definición, el método tiene un límite máximo de 5000 mutaciones aceptadas. Este límite representa un valor muy alto donde se considera improductivo seguir

intentando mutaciones porque probablemente no sea posible encontrar el resultado deseado. Es posible modificar este valor para una ejecución en particular mediante el parámetro `maxiterations`, de la forma:

```
--maxiterations [número máximo de iteraciones]
```

Por ejemplo:

```
--maxiterations 30      Finaliza la ejecución si se alcanzan 30 mutaciones
```

El objetivo de la modificación puede ser disminuir el número máximo de mutaciones, para obtener un resultado más rápidamente, aunque de menor calidad. También se puede incrementar el número máximo de mutaciones para linkers especialmente largos o con restricciones muy fuertes para el diseño. En todos los casos donde se alcanza el límite de mutaciones se informa lo sucedido y la secuencia correspondiente a esa iteración.

4.2.6. Formato y detalles del resultado

Por definición el método sólo devuelve el resultado final de la ejecución informando, además, la condición de finalización. Es posible extender o modificar el formato de salida mediante las siguientes opciones:

```
--logoutput [logpath]
```

Guarda el historial de mutaciones en un archivo .log ubicado en logpath.

```
--verbose
```

Devuelve por pantalla una salida detallada de la ejecución.

```
--stepped
```

Ejecución paso a paso: espera el input del usuario en cada paso del método.

Capítulo 5

Evaluaciones y análisis de resultados

5.1. Parámetros de ejecución

Todas las evaluaciones se realizaron con los parámetros estándar del método descritos en el manual (ver capítulo 4). El equipamiento y software utilizado para las pruebas es el siguiente:

Sistema operativo: Ubuntu 14.04 - Kernel version: 3.13.0

CPU: Intel(R) Core(TM) i7-4770 CPU @ 3.40GHz

Memoria RAM: 16GB

5.2. Estimación del parámetro β

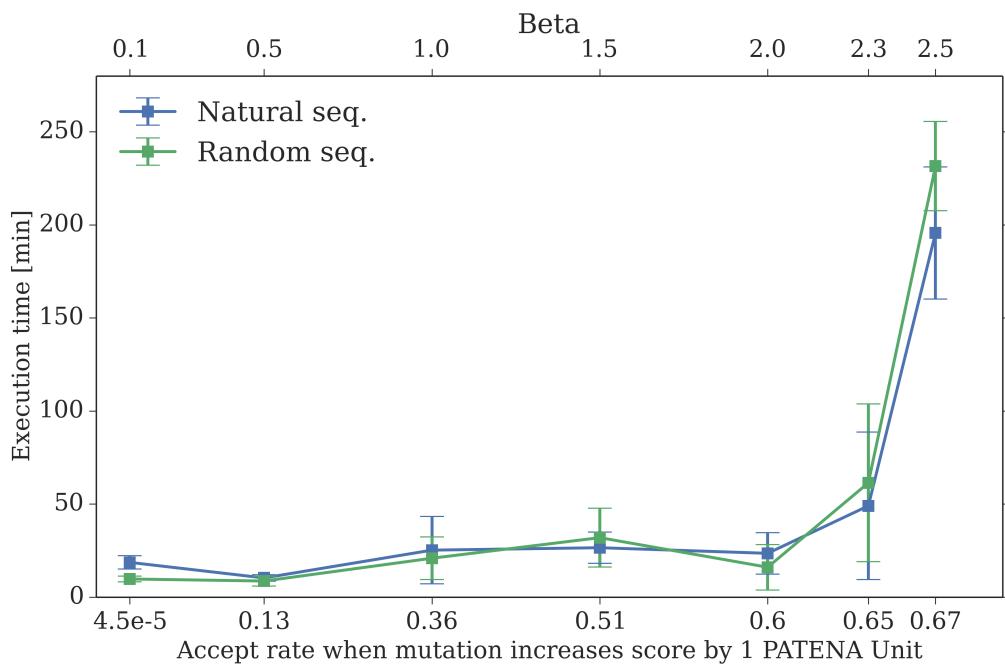
En las secciones previas hemos propuesto un método de diseño que se basa en búsqueda estocástica sobre el espacio de secuencias. El método de búsqueda es guiado por el resultado de una función que asigna un puntaje a cada secuencia, el cual resulta siempre mayor o igual a 0. La secuencia resultante buscada se caracteriza por tener un puntaje igual a 0. Por lo tanto, el método puede verse como la búsqueda de un mínimo global sobre la superficie que relaciona cada posible secuencia con su valor de puntaje.

El comportamiento del método de búsqueda depende de un único parámetro β cuyo valor (o rango) óptimo está fuertemente asociado a las características de esta superficie. Un valor de β más grande tiene un porcentaje de aceptación mayor de

mutaciones que aumentan el puntaje de la secuencia. Por lo tanto, permite la exploración de superficies que requieren superar mínimos locales para alcanzar secuencias con puntaje 0. Un valor de β más chico implica recorrer un camino más directo hacia el mínimo ya que se aceptan menos mutaciones que aumenten el puntaje, pero esto podría requerir la evaluación de una gran cantidad de posibilidades. Incluso, si la búsqueda se estanca en un mínimo local, un valor muy chico de β solo podría encontrar la solución en un tiempo muy grande. Como vimos en el capítulo 2, la decisión de aceptación está basada en el uso de la ecuación 2.1, la cual siempre devuelve un valor > 0 para cualquier β positivo. Por lo tanto, la probabilidad de aceptación nunca es 0 y no es imposible encontrar el resultado si es que existe, aún cuando el tiempo que se demore sea demasiado largo para un uso práctico del método.

A pesar que, previo a desarrollar la aplicación asumimos que el espacio de soluciones era considerablemente grande, no sabemos cómo es exactamente la superficie que asocia cada secuencia con el puntaje definido por las evaluaciones. De esta forma, inicialmente no tenemos ningún conocimiento de cómo se relaciona el valor del parámetro β con la ejecución resultante, ni cuáles serán los valores de β que nos permitirán obtener diseños de la forma más eficiente. La forma más directa de conocer cuál es el rango (o valor) óptimo de β es mediante la evaluación del tiempo de ejecución, es decir, el tiempo total requerido para la búsqueda del diseño resultante. Para esto, se midió el tiempo de ejecución para distintas corridas que utilizan valores de β en el rango 0.1-2.5. Puntualmente se analizan los valores en el conjunto (0.1, 0.5, 1.0, 1.5, 2.0, 2.3, 2.5). Se realizaron 6 ejecuciones para cada valor de β analizado, de las cuales 3 se realizan a partir de secuencias definidas (obtenidas de proteínas naturales) y otras 3 a partir de secuencias generadas aleatoriamente. En todos los casos con una longitud de 50 residuos.

En la figura 5.1 se muestran los tiempos medios asociados a cada conjunto de evaluaciones. En primer lugar vemos que no hay una diferencia significativa constante entre las ejecuciones que inician a partir de secuencias naturales y las que lo hacen a partir de secuencias aleatorias. Por otro lado, se ve que el tiempo de ejecución es altamente variable para la mayoría de los valores de β . Esta variabilidad, que se basa en las propiedades estocásticas de la búsqueda, no permite definir un valor óptimo puntual. Se puede ver, igualmente, que el tiempo de ejecución es significativamente mayor en el caso de β muy grande (2.3 y 2.5), indicando que el rango óptimo de valores está ubicado hacia el extremo inferior. Aunque el rango completo 0.1-2.0 es aceptable, definimos a $\beta = 1.0$ como un valor que nos permitirá realizar la ejecución en un tiempo aceptable, quedando como valor estándar de la herramienta.

Figura 5.1: Dependencia del tiempo de ejecución con β

5.3. Análisis detallado de la ejecución

El rango de β obtenido es útil para poder encontrar resultados eficientemente. No obstante, si queremos conocer más en detalle las propiedades de la búsqueda, hace falta desglosar la ejecución. Los detalles de las búsquedas permiten, además, inferir ciertas características de la superficie que se está explorando. Sabiendo que las propiedades de la búsqueda son, puntualmente, la cantidad total de mutaciones aceptadas y la cantidad de mutaciones que se intentan para cada ejecución, en lo que resta de esta sección analizaremos cómo cambian los perfiles de estas variables según el valor de β .

Para tener una primera idea de esta dependencia, analizaremos los perfiles de ejecución para dos valores puntuales de β : $\beta = 0.5$ y $\beta = 2.4$. En el caso del β más chico (0.5), la probabilidad de aceptar una mutación para un incremento de 1 en el puntaje es de 13 %, mientras que para el β más grande (2.4) esta probabilidad es de 65,9 %. Por lo tanto, los dos valores podrían clasificarse como cercanos a los extremos dentro del esquema de decisión basado en el puntaje. Para cada valor de β se realizaron 6 corridas independientes. La mitad de estas corridas se iniciaron a partir de secuencias generadas aleatoriamente, mientras que las restantes se iniciaron a partir de secuencias naturales definidas (distintas entre si). En todos los casos la longitud de la secuencia fue de 30 aminoácidos. Los resultados se muestran en las figuras 5.2 y 5.3.

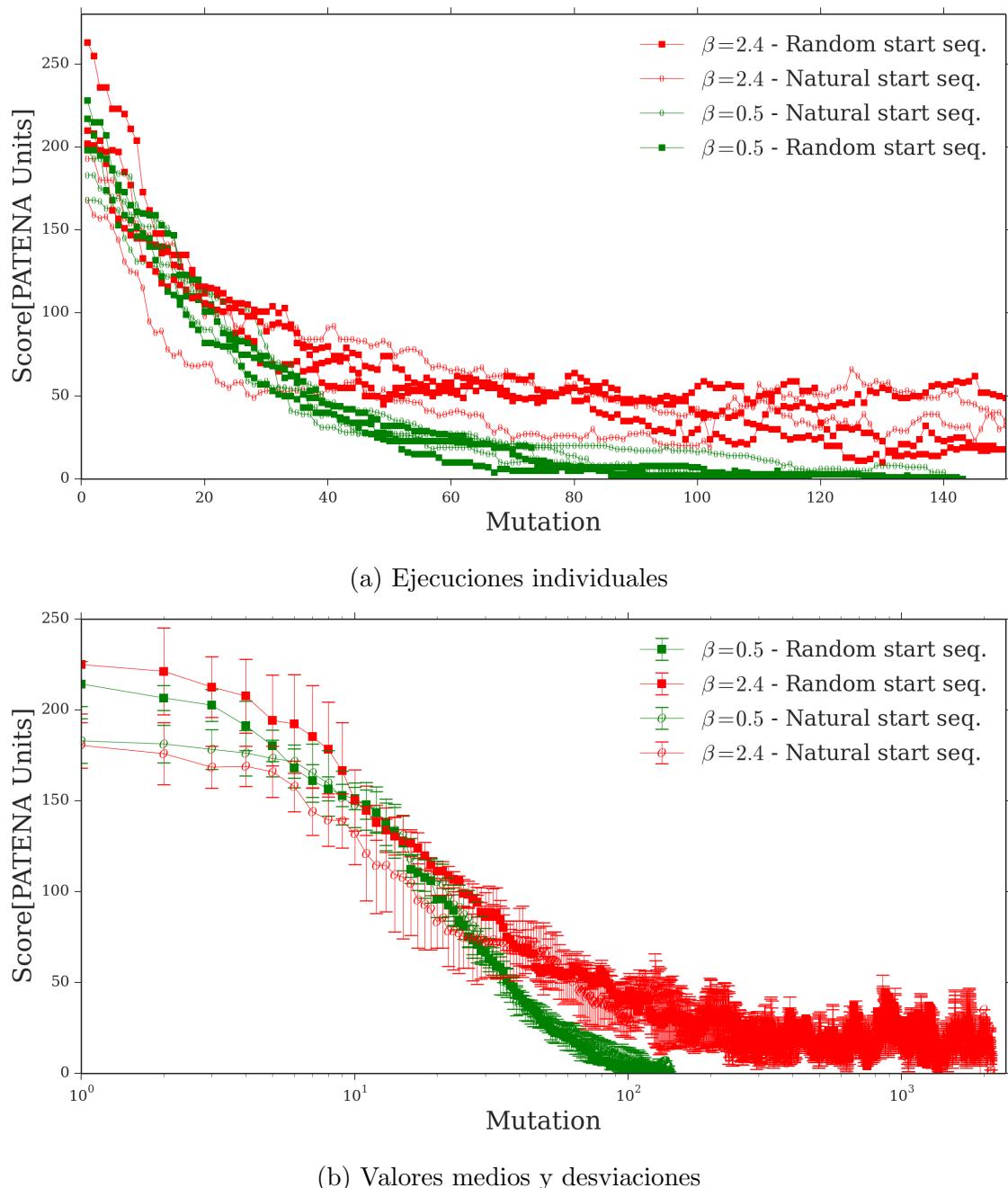
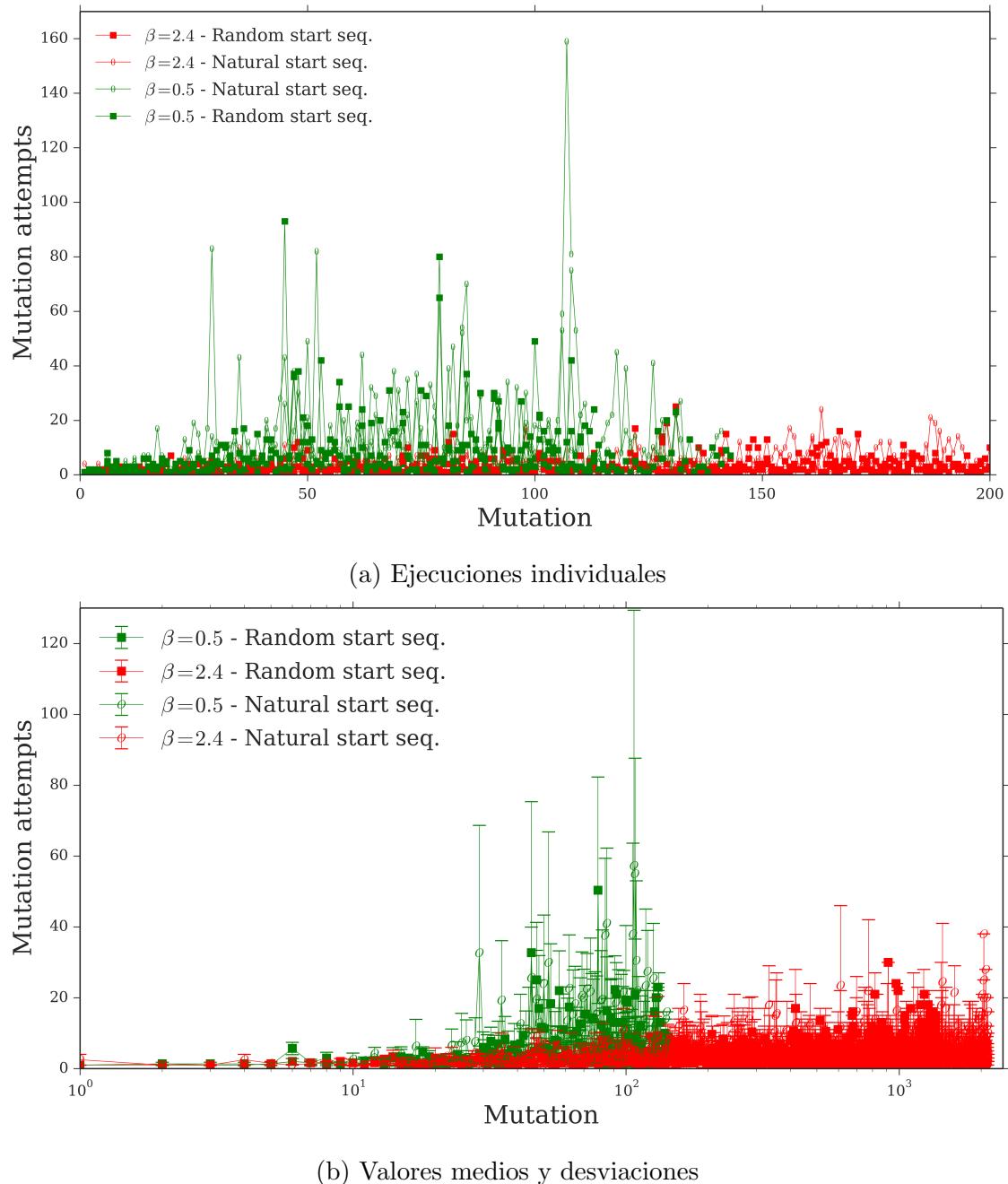


Figura 5.2: Perfil de puntajes asociado a cada iteración para distintos β

Figura 5.3: Dependencia del número de intentos de mutación con β

En el gráfico 5.2 se muestra la dependencia del puntaje con las mutaciones aplicadas, tanto los valores medios (y desviaciones estándar) para todas las ejecuciones mencionadas (gráfico 5.2b), como el detalle de las ejecuciones individuales (gráfico 5.2a). El perfil que se muestra permite aclarar mejor los conceptos mencionados previamente acerca de la dependencia de β con el comportamiento de la búsqueda. Un valor más grande de β realiza una mayor exploración del espacio de soluciones, lo que se ve reflejado en un rango mucho más amplio de mutaciones aplicadas para alcanzar el resultado. Se ve claramente que hay una diferencia en el orden de magnitud de la cantidad de mutaciones aplicadas, donde las ejecuciones con $\beta = 2.4$ pueden llegar a alcanzar más de 2000. Por su parte, el valor más chico de β permite alcanzar el objetivo en un número mucho menor de mutaciones, lo cual parece indicar que hay un camino que conduce desde cada punto de inicio hacia un resultado bajando continuamente el valor del puntaje, es decir, sin atravesar ninguna barrera. De todas formas, a partir de este experimento no podemos extraer datos concretos sobre las características de este recorrido hacia el mínimo.

En el gráfico 5.3 se muestra el número de intentos requeridos para lograr la aceptación de cada mutación aplicada a lo largo de la ejecución. Nuevamente se muestran los valores medios (y sus desviaciones correspondientes) en el gráfico 5.3b, y el detalle de las ejecuciones individuales (gráfico 5.3a). Vemos que el reducido número de mutaciones necesarias que resulta de un valor bajo de β se produce a costa de una cantidad más elevada de intentos previos. Es decir, se evalúa una mayor cantidad de posibles mutaciones hasta que eventualmente una sea aceptada, lo que representa la dificultad del camino hacia el valor mínimo buscado. Esta condición se incrementa en las mutaciones cercanas al fin de la ejecución, que se corresponden con los valores más bajos de puntaje (cercaos a 0).

Comprobamos, entonces, que un valor más grande de β implica un incremento considerable en el número de mutaciones requeridas, aunque un valor menor de β tendrá una mayor cantidad de intentos de mutación por iteración. Sabiendo que ambas propiedades impactan negativamente en el proceso de búsqueda incrementando el tiempo de ejecución, intentaremos ver ahora cómo, en el rango efectivo de β , se crea el balance óptimo entre estos parámetros. Para esto, analizaremos el perfil completo de las ejecuciones realizadas en la sección previa para encontrar el rango óptimo de β , lo cual se muestra en las figuras 5.4a y 5.4b (no se realiza la separación entre ejecuciones iniciadas a partir de secuencias naturales o secuencias random).

Lo que vemos en la figura 5.4a es que el número de intentos de mutación por iteración baja al aumentar el valor de β , lo cual es esperable de acuerdo a lo visto previamente. Sin embargo, según se puede ver en el gráfico 5.4b el número total de

intentos de mutación no siempre tiene una disminución correspondiente. En algunas regiones, principalmente para los valores de β más altos, el aumento del número de iteraciones es mucho más significativo que la disminución en los intentos de mutación de cada iteración. Esto se debe a que, para valores bajos de β , los intentos de mutación son considerablemente bajos, pero también se debe a que el incremento en el número de iteraciones requeridas es muy alto para valores grandes de β . El resultado es que, para valores de β superiores a 2.0, el incremento en el número de mutaciones requeridas es tan grande que produce, también, un incremento en el número total de intentos de mutación. Esto rompe el balance entre ambas propiedades y genera un aumento del tiempo de ejecución, como se vió en la sección anterior.

5.4. Dependencia con la longitud de la secuencia

Una vez definido un valor de β que permite un balance estable entre intentos de mutación y mutaciones aceptadas, queremos saber cómo se traduce esto en un tiempo de ejecución concreto para distintos casos de uso de la herramienta. Se realizan, entonces, evaluaciones del tiempo de ejecución en función de la longitud de la secuencia, lo que nos dará una idea del orden de tiempo que demoran las ejecuciones y como éste varía con la longitud. En el gráfico 5.5 se muestran los resultados de distintas ejecuciones individuales de la aplicación, utilizando siempre el valor de β definido previamente (1.0) pero con longitudes de secuencia variable, tanto para secuencias aleatorias como para secuencias iniciales naturales.

Si bien los resultados parecen indicar una relación aproximadamente lineal entre la longitud de la secuencia y el tiempo de ejecución, la relevancia de estos resultados es que nos permiten saber que la solución es aplicable a secuencias en el rango esperado de utilización de la herramienta.

5.5. Análisis de diseños resultantes

Ahora que tenemos fijados todos los aspectos de ejecución del método y que efectivamente podemos obtener, en un tiempo aceptable, secuencias que cumplen con requerimientos definidos, nos centraremos en conocer más acerca de los diseños resultantes.

Por definición del método, sabemos que los resultados tendrán las propiedades positivas buscadas y no tendrán las características negativas pero, hasta el momento, no sabemos nada más acerca de los diseños que se pueden obtener. Para comenzar a analizar los diseños resultantes se realizaron un total de 74 ejecuciones independien-

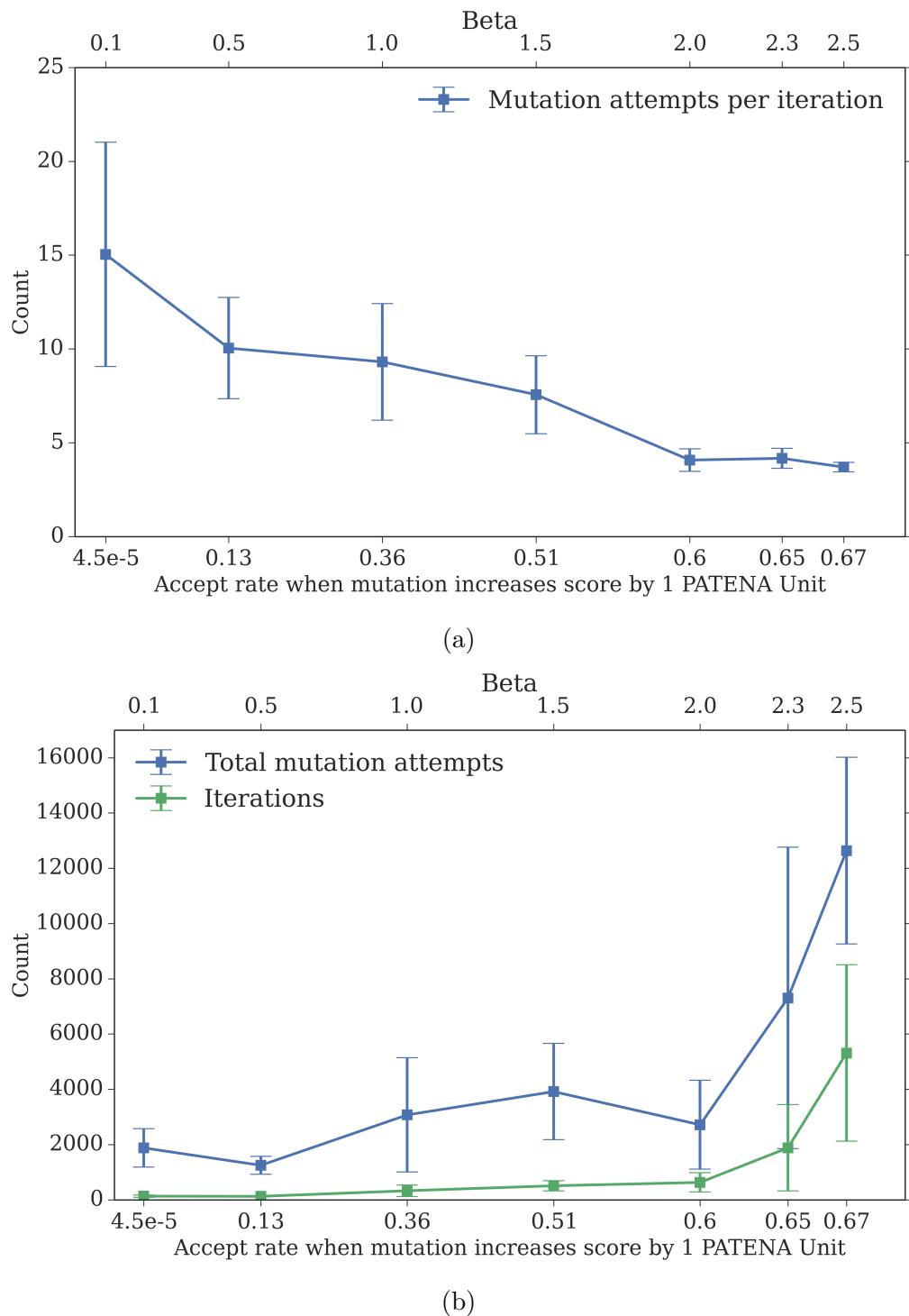


Figura 5.4: Dependencia de los parámetros de la búsqueda con el valor de beta

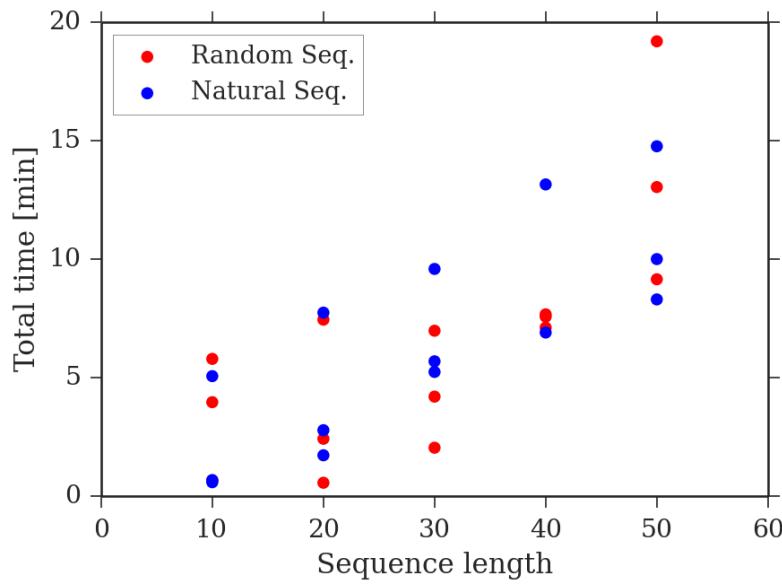


Figura 5.5: Dependencia del tiempo de ejecución con la longitud de la secuencia

tes a partir de una misma secuencia inicial de largo=30 (MALWMRLLLPLLALLALWGPDPAAAFVNQHL). De forma independiente, mediante la herramienta RandSeq [11], se obtuvieron 1000 secuencias al azar de largo=30 utilizando la composición estándar de nuestro método. Este *background* representa la máxima diversidad secuencial alcanzable con esta composición. Dado que todas las secuencias random, resultantes e inicial tienen la misma longitud, las evaluaciones de identidad entre estas secuencias (mostradas en la figura 5.6) se realizaron a partir del número de residuos idénticos encontrados en posiciones correspondientes del par de secuencias. El porcentaje de identidad resultante se calcula como:

$$\left(\frac{\text{Número de posiciones con residuos idénticos}}{\text{Largo de la secuencia}} \right) * 100 \quad (5.1)$$

Como se ve en la figura 5.6 (izquierda), los diseños obtenidos a partir de este ensayo, a pesar de tener una similitud mayor que la encontrada entre secuencias random, se muestran como un conjunto considerablemente heterogéneo. Esto confirma que el método permite obtener la diversidad buscada en los diseños resultantes a partir del uso de la composición estándar. Por otro lado, en la figura 5.6 (derecha) se muestra que el conjunto de resultados, en comparación con el conjunto de secuencias random, tiene un porcentaje mayor de similitud con la secuencia inicial. De esta forma, aún habiendo una gran cantidad de mutaciones, los diseños obtenidos conservan cierta similitud con la secuencia inicial propuesta. Esta similitud remanente

comprende una ventaja del método ya que permite al usuario proponer un diseño y obtener resultados que presentan cierta similitud con este.

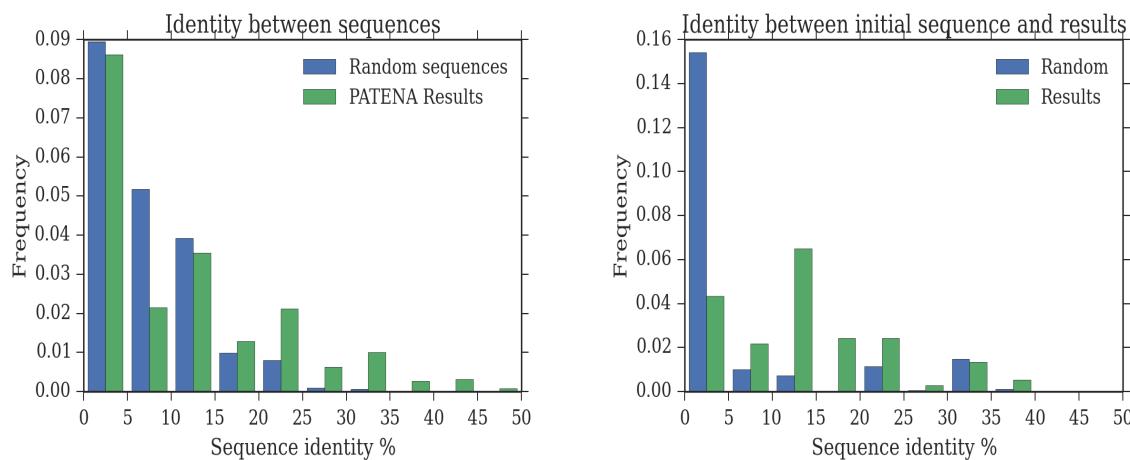


Figura 5.6: Histograma de identidad entre las secuencias resultantes entre si (izquierda) y entre éstas y la secuencia inicial (derecha)

Sin embargo, este análisis de identidad global entre secuencias no permite conocer si la similitud se encuentra localizada en alguna posición específica. La extracción de un logo secuencial [94] permite obtener, a través de una representación gráfica, un análisis detallado de la similitud dentro de un conjunto de secuencias. En la figura 5.7 se muestra el logo obtenido a partir de los resultados correspondientes a las 74 corridas ejecutadas.

Como se ve, la similitud entre los resultados se concentra en las posiciones que inicialmente poseían residuos de Glicina y Prolina (posiciones 9,18,19,21), y los residuos más representados en estas posiciones son, justamente, los mismos que se encontraban en la secuencia inicialmente. Como vimos en el capítulo 1, estos son los residuos más encontrados en secuencias linkers tanto naturales como artificiales debido a sus propiedades fisicoquímicas, que permiten caracterizarlos como residuos que favorecen el desorden intrínseco en la conformación del polipéptido. En estos casos, los residuos iniciales correspondientes a estas posiciones están más conservados en el diseño final. Cabe aclarar que el grado de conservación encontrado es relativamente bajo ya que para secuencias de proteínas, como se describe en [30], la conservación toma valores entre 0 y $\log_2 20 \approx 4,32$. Por otro lado, el resto de las posiciones mantienen una gran diversidad que reflejan la heterogeneidad mostrada en las figuras previas.

Por último queremos saber ahora si, además de la diversidad resultante que mostramos hasta aquí, estamos efectivamente minimizando el costo metabólico de

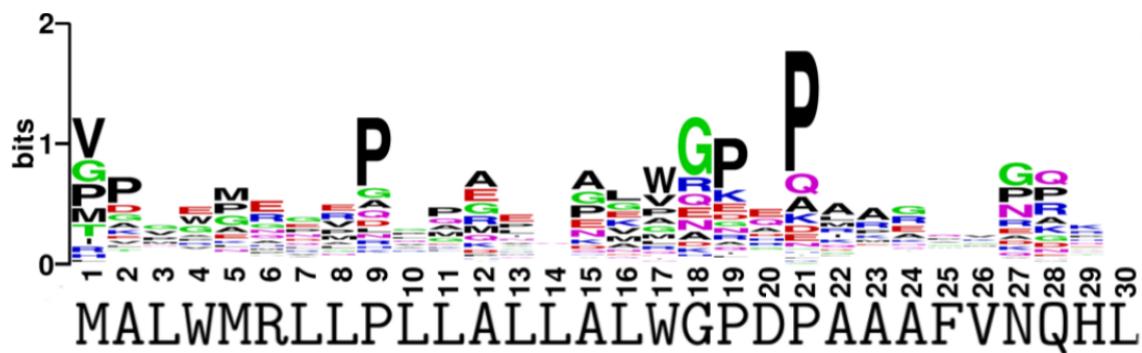


Figura 5.7: Representación gráfica de la similitud secuencial entre los resultados. Debajo del logo se muestra la secuencia inicial a partir de la cual se obtuvieron los resultados. Figura obtenida utilizando [30]

los diseños obtenidos. Es importante comprobar esto ya que el objetivo de utilizar la composición estándar extraída de Swissprot (detallado en 2.2.1) era obtener un correcto balance entre diversidad y costo metabólico asociado. Para evaluarlo comparamos la frecuencia de cada aminoácido, extraída de los 74 resultados obtenidos, con la frecuencia esperada de acuerdo a la composición estándar que utilizamos en el método. Los resultados, mostrados en el gráfico 5.8, muestran una correspondencia entre la frecuencia esperada y la resultante en este pequeño conjunto de resultados. El caso que más se desvía de la correspondencia es el aminoácido Prolina, donde la frecuencia observada es considerablemente mayor que la esperada, lo cual es entendible si tenemos en cuenta las propiedades de este residuo comentadas previamente. Esta desviación es correspondiente, además, con la conservación encontrada en los resultados que se vió en la figura 5.7.

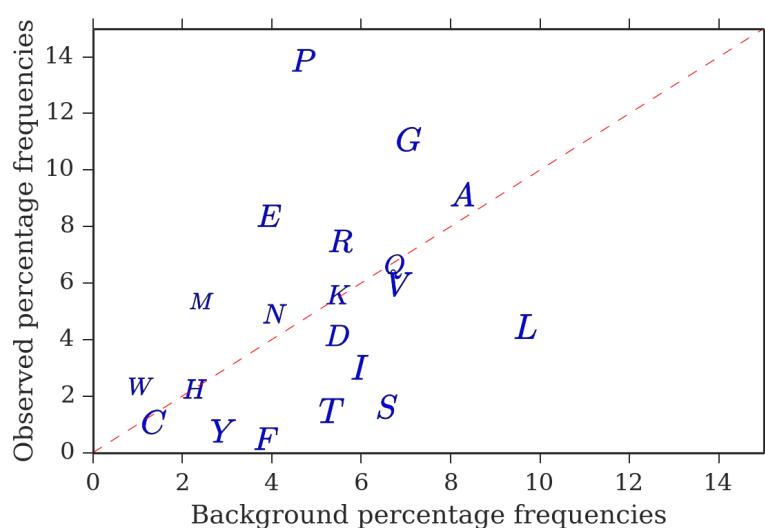


Figura 5.8: Comparación de frecuencias esperadas y frecuencias observadas en los diseños resultantes

Capítulo 6

Conclusiones y trabajo a futuro

- Hemos implementado un nuevo método que permite obtener secuencias linker flexibles, sin estructura residual, sin interacciones conocidas con otros componentes celulares, sin secuencias repetitivas difíciles de clonar y expresar y con un costo metabólico por aminoácido similar al de las proteínas naturales.
- El método permite obtener secuencias linker en un tiempo de minutos, adecuado para el trabajo de diseño de proteínas multidominio. Los diseños resultantes son diversos en secuencia, aun partiendo de la misma secuencia inicial. Estos dos resultados son compatibles con la hipótesis inicial de que existe un gran número de secuencias con las características que definimos como deseables para un linker.
- El método desarrollado se provee en forma estandarizada y evaluada incluyendo, además, diversas funcionalidades que permiten personalizar la ejecución. De esta forma, se da la posibilidad al usuario de obtener resultados específicos de acuerdo al contexto de aplicación que tendrá el diseño resultante.
- En el futuro inmediato se espera poder desarrollar un servidor web que provea la posibilidad de obtener fácilmente secuencias linker. Este paso implica la implementación de una interfaz simple para que pueda ser usada por cualquier usuario experimental, adaptando las funcionalidades del método desarrollado.

Bibliografía

- [1] Amino acid composition (%) in the UniProtKB/Swiss-Prot data bank. <http://web.expasy.org/protscale/pscale/A.A.Swiss-Prot.html>. [Release notes for UniProtKB/SwissProt release April 2013].
- [2] Biopython. http://biopython.org/wiki/Main_Page. [Online; accessed 19-July-2015].
- [3] ELM WebServer. <http://elm.eu.org/>. [Online; accessed 19-July-2015].
- [4] Expresiones regulares(web). <http://www.regular-expressions.info/>. [Online; accessed 19-July-2015].
- [5] Git. <https://git-scm.com/book/en/v2/Getting-Started-Installing-Git>. [Online; accessed 19-July-2015].
- [6] IUPred Download link. <http://iupred.enzim.hu/Downloads.php>, . [Online; accessed 19-July-2015].
- [7] IUPred Web server. <http://iupred.enzim.hu/>, . [Online; accessed 19-July-2015].
- [8] Perl. <https://www.perl.org/get.html>. [Online; accessed 19-July-2015].
- [9] PROSITE web. <http://prosite.expasy.org/>. [Online; accessed 19-July-2015].
- [10] Python. <https://www.python.org/downloads/>. [Online; accessed 19-July-2015].
- [11] RandSeq. <http://web.expasy.org/randseq/>. [Online; accessed 19-July-2015].
- [12] ScanProsite web. <http://prosite.expasy.org/scanprosite/>. [Online; accessed 19-July-2015].

- [13] Standalone BLAST Setup for Unix. <http://www.ncbi.nlm.nih.gov/books/NBK52640/>, . [Online; accessed 19-July-2015].
- [14] TANGO WebServer. <http://tango.crg.es/>. [Online; accessed 19-July-2015].
- [15] TMHMM v2.0. <http://www.cbs.dtu.dk/services/TMHMM/>. [Online; accessed 19-October-2015].
- [16] Waltz Web Server. <http://waltz.switchlab.org/>. [Online; accessed 19-July-2015].
- [17] WEB - BLAST: Basic local alignment search tool. <http://blast.ncbi.nlm.nih.gov/Blast.cgi>, . [Online; accessed 19-July-2015].
- [18] Neeraj J Agrawal, Sandeep Kumar, Xiaoling Wang, Bernhard Helk, Satish K Singh, y Bernhardt L Trout. Aggregation in protein-based biotherapeutics: Computational studies and tools to identify aggregation-prone regions. *Journal of pharmaceutical sciences*, 100(12):5081–5095, 2011.
- [19] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, y David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.
- [20] Ryoichi Arai, Hiroshi Ueda, Atsushi Kitayama, Noriho Kamiya, y Teruyuki Nagamune. Design of the linkers which effectively separate domains of a bifunctional fusion protein. *Protein engineering*, 14(8):529–532, 2001.
- [21] Ryoichi Arai, Willy Wriggers, Yukihiro Nishikawa, Teruyuki Nagamune, y Tetsuro Fujisawa. Conformations of variably linked chimeric proteins evaluated by synchrotron X-ray small-angle scattering. *PROTEINS: Structure, Function, and Bioinformatics*, 57(4):829–838, 2004.
- [22] Patrick Argos. An investigation of oligopeptides linking domains in protein tertiary structures and possible candidates for general gene fusion. *Journal of molecular biology*, 211(4):943–958, 1990.
- [23] Amos Bairoch y Rolf Apweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic acids research*, 28(1):45–48, 2000.
- [24] Robert E Bird, Karl D Hardman, James W Jacobson, Syd Johnson, Bennett M Kaufman, Shwu-Maan Lee, Timothy Lee, Sharon H Pope, Gary S Riordan, y

- Marc Whitlow. Single-chain antigen-binding proteins. *Science*, 242(4877):423–426, 1988.
- [25] Jessica Walton Chen, Pedro Romero, Vladimir N Uversky, y A Keith Dunker. Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions. *Journal of proteome research*, 5(4):879–887, 2006.
- [26] Jessica Walton Chen, Pedro Romero, Vladimir N Uversky, y A Keith Dunker. Conservation of intrinsic disorder in protein domains and families: II. functions of conserved disorder. *Journal of proteome research*, 5(4):888–898, 2006.
- [27] Xiaoying Chen, Jennica L Zaro, y Wei-Chiang Shen. Fusion protein linkers: property, design and functionality. *Advanced drug delivery reviews*, 65(10):1357–1369, 2013.
- [28] Chiquito J Crasto y Jin-an Feng. LINKER: a program to generate linker sequences for fusion proteins. *Protein engineering*, 13(5):309–312, 2000.
- [29] Donna E Crone, Christian Schenkelberg, Christopher Bystroff, Derek J Pittman, Keith Fraser, Stephen Macari, y Yao-Ming Huang. *GFP-based biosensors*. INTECH Open Access Publisher, 2013.
- [30] Gavin E Crooks, Gary Hon, John-Marc Chandonia, y Steven E Brenner. Weblogo: a sequence logo generator. *Genome research*, 14(6):1188–1190, 2004.
- [31] Norman E Davey, Kim Van Roey, Robert J Weatheritt, Grischa Toedt, Bo-ra Uyar, Brigitte Altenberg, Aidan Budd, Francesca Diella, Holger Dinkel, y Toby J Gibson. Attributes of short linear motifs. *Molecular BioSystems*, 8(1):268–281, 2012.
- [32] Mercedes Kuroski de Bold, William P Sheffield, Amy Martinuk, Varsha Bhakta, Louise Eltringham-Smith, y J Adolfo. Characterization of a long-acting recombinant human serum albumin-atrial natriuretic factor (ANF) expressed in *Pichia pastoris*. *Regulatory peptides*, 175(1):7–10, 2012.
- [33] Edouard De Castro, Christian JA Sigrist, Alexandre Gattiker, Virginie Bulliard, Petra S Langendijk-Genevaux, Elisabeth Gasteiger, Amos Bairoch, y Nicolas Hulo. ScanPosite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic acids research*, 34(suppl 2):W362–W365, 2006.

- [34] Manuela López de la Paz, Kenneth Goldie, Jesús Zurdo, Emmanuel Lacroix, Christopher M Dobson, Andreas Hoenger, y Luis Serrano. De novo designed peptide-based amyloid fibrils. *Proceedings of the National Academy of Sciences*, 99(25):16052–16057, 2002.
- [35] Manuela Lopez de la Paz y Luis Serrano. Sequence determinants of amyloid fibril formation. *Proceedings of the National Academy of Sciences*, 101(1):87–92, 2004.
- [36] Paul Dean. Functional domains and motifs of bacterial type III effector proteins and their roles in infection. *FEMS microbiology reviews*, 35(6):1100–1125, 2011.
- [37] Frederick A Dick y Seth M Rubin. Molecular mechanisms underlying RB protein function. *Nature reviews Molecular cell biology*, 14(5):297–306, 2013.
- [38] Holger Dinkel, Kim Van Roey, Sushama Michael, Norman E Davey, Robert J Weatheritt, Diana Born, Tobias Speck, Daniel Krüger, Gleb Grebnev, Marta Kubań, et al. The eukaryotic linear motif resource ELM: 10 years and counting. *Nucleic acids research*, pág. gkt1047, 2013.
- [39] Zsuzsanna Dosztányi, Veronika Csizmok, Peter Tompa, y István Simon. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, 21(16):3433–3434, 2005.
- [40] Zsuzsanna Dosztanyi, Veronika Csizmok, Peter Tompa, y Istvan Simon. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *Journal of molecular biology*, 347(4):827–839, 2005.
- [41] Zsuzsanna Dosztányi, Bálint Mészáros, y István Simon. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics*, 25(20):2745–2746, 2009.
- [42] H Jane Dyson y Peter E Wright. Intrinsically unstructured proteins and their functions. *Nature reviews Molecular cell biology*, 6(3):197–208, 2005.
- [43] David Eliezer. Biophysical characterization of intrinsically disordered proteins. *Current opinion in structural biology*, 19(1):23–30, 2009.

- [44] Toon H Evers, Elisabeth MWM van Dongen, Alex C Faesen, EW Meijer, y Maarten Merkx. Quantitative understanding of the energy transfer between fluorescent proteins connected via flexible peptide linkers. *Biochemistry*, 45(44):13183–13192, 2006.
- [45] Zhanmin Fan, Joshua R Werkman, y Ling Yuan. Engineering of a multifunctional hemicellulase. *Biotechnology letters*, 31(5):751–757, 2009.
- [46] Marcus Fändrich y Christopher M Dobson. The behaviour of polyamino acids reveals an inverse side chain effect in amyloid structure formation. *The EMBO journal*, 21(21):5682–5690, 2002.
- [47] Ana-Maria Fernandez-Escamilla, Frederic Rousseau, Joost Schymkowitz, y Luis Serrano. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature biotechnology*, 22(10):1302–1306, 2004.
- [48] Douglas M Fowler, Atanas V Koulov, William E Balch, y Jeffery W Kelly. Functional amyloid—from bacteria to humans. *Trends in biochemical sciences*, 32(5):217–224, 2007.
- [49] Paul E Fraser, DR McLachlan, WK Surewicz, CA Mizzen, AD Snow, JT Nguyen, y DA Kirschner. Conformation and fibrillogenesis of alzheimer $\alpha\beta$ peptides with selected substitution of charged residues. *Journal of molecular biology*, 244(1):64–73, 1994.
- [50] Monika Fuxreiter, Peter Tompa, y István Simon. Local structural disorder imparts plasticity on linear motifs. *Bioinformatics*, 23(8):950–956, 2007.
- [51] Richard A George y Jaap Heringa. An analysis of protein domain linkers: their classification and role in protein folding. *Protein Engineering*, 15(11):871–879, 2002.
- [52] Bernard R Glick. Metabolic load and heterologous gene expression. *Biotechnology advances*, 13(2):247–261, 1995.
- [53] Kannan Gunasekaran, Chung-Jung Tsai, Sandeep Kumar, David Zanuy, y Ruth Nussinov. Extended disordered proteins: targeting function with less scaffold. *Trends in biochemical sciences*, 28(2):81–85, 2003.
- [54] Johnny Habchi, Peter Tompa, Sonia Longhi, y Vladimir N Uversky. Introducing protein intrinsic disorder. *Chemical reviews*, 114(13):6561–6588, 2014.

- [55] Stavros J Hamodrakas. Protein aggregation and amyloid fibril formation prediction software from primary sequence: towards controlling the formation of bacterial inclusion bodies. *Febs Journal*, 278(14):2428–2435, 2011.
- [56] Caroline Hilbich, Brigitte Kisters-Woike, Jennifer Reed, Colin L Masters, y Konrad Beyreuther. Substitutions of hydrophobic amino acids reduce the amyloidogenicity of alzheimer's disease β a4 peptides. *Journal of molecular biology*, 228(2):460–473, 1992.
- [57] Christine A Hrycyna, Lisa E Airan, Ursula A Germann, Suresh V Ambudkar, Ira Pastan, y Michael M Gottesman. Structural flexibility of the linker region of human P-glycoprotein permits ATP hydrolysis and drug transport. *Biochemistry*, 37(39):13660–13673, 1998.
- [58] IBIU. Linker database web. <http://www.ibi.vu.nl/programs/linkerdbwww/>. [Online; accessed 19-July-2015].
- [59] Mitsuo Ikebe, Taketoshi Kambara, Walter F Stafford, Masataka Sata, Eisaku Katayama, y Reiko Ikebe. A hinge at the central helix of the regulatory light chain of myosin is critical for phosphorylation-dependent regulation of smooth muscle myosin motor activity. *Journal of Biological Chemistry*, 273(28):17702–17707, 1998.
- [60] Masahiro Iwakura y Tsutomu Nakamura. Effects of the length of a glycine linker connecting the N-and C-termini of a circularly permuted dihydrofolate reductase. *Protein engineering*, 11(8):707–713, 1998.
- [61] Martin Karplus, J Andrew McCammon, y Warner L Peticolas. The internal dynamics of globular protein. *CRC critical reviews in biochemistry*, 9(4):293–349, 1981.
- [62] Teresa Krick, Nina Verstraete, Leonardo G Alonso, David A Shub, Diego U Ferreiro, Michael Shub, y Ignacio E Sánchez. Amino acid metabolism conflicts with protein diversity. *Molecular biology and evolution*, 31(11):2905–2912, 2014.
- [63] Anders Krogh, BjoÈrn Larsson, Gunnar Von Heijne, y Erik LL Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of molecular biology*, 305(3):567–580, 2001.

- [64] Emmanuel Lacroix, Ana Rosa Viguera, y Luis Serrano. Elucidating the folding problem of α -helices: local motifs, long-range electrostatics, ionic-strength dependence and prediction of nmr parameters. *Journal of molecular biology*, 284(1):173–191, 1998.
- [65] John H Laity, H Jane Dyson, y Peter E Wright. DNA-induced α -helix capping in conserved linker sequences is a determinant of binding affinity in Cys 2-His 2 zinc fingers. *Journal of molecular biology*, 295(4):719–727, 2000.
- [66] Mario Lebendiker y Tsafi Danieli. Production of prone-to-aggregate proteins. *FEBS letters*, 588(2):236–246, 2014.
- [67] Rune Linding, Joost Schymkowitz, Frederic Rousseau, Francesca Diella, y Luis Serrano. A comparative study of the relationship between protein structure and β -aggregation in globular and intrinsically disordered proteins. *Journal of molecular biology*, 342(1):345–353, 2004.
- [68] Chengcheng Liu, Ju Xin Chin, y Dong-Yup Lee. SynLinker: an integrated system for designing linkers and synthetic fusion proteins. *Bioinformatics*, pág. btv447, 2015.
- [69] Peter Ljungcrantz, Helen Carlsson, Mats Olle Mansson, Peter Buckel, Klaus Mosbach, y Leif Buelow. Construction of an artificial bifunctional enzyme, beta-galactosidase/galactose dehydrogenase, exhibiting efficient galactose channeling. *Biochemistry*, 28(22):8786–8792, 1989.
- [70] Thorsten Lührs, Christiane Ritter, Marc Adrian, Dominique Riek-Loher, Bernd Bohrmann, Heinz Döbeli, David Schubert, y Roland Riek. 3D structure of Alzheimer's amyloid- β (1–42) fibrils. *Proceedings of the National Academy of Sciences of the United States of America*, 102(48):17342–17347, 2005.
- [71] Dahai Luo, Na Wei, Danny N Doan, Prasad N Paradkar, Yuwen Chong, Andrew D Davidson, Masayo Kotaka, Julien Lescar, y Subhash G Vasudevan. Flexibility between the protease and helicase domains of the dengue virus NS3 protein conferred by the linker region and its functional implications. *Journal of biological chemistry*, 285(24):18817–18827, 2010.
- [72] O Sumner Makin, Edward Atkins, Paweł Sikorski, Jan Johansson, y Louise C Serpell. Molecular basis for amyloid fibril formation and stability. *Proceedings of the National Academy of Sciences of the United States of America*, 102(2):315–320, 2005.

- [73] Albert H Mao, Scott L Crick, Andreas Vitalis, Caitlin L Chicoine, y Rohit V Pappu. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proceedings of the National Academy of Sciences*, 107(18):8183–8188, 2010.
- [74] Sebastian Maurer-Stroh, Maja Debulpaepe, Nico Kuemmerer, Manuela Lopez de la Paz, Ivo Cristiano Martins, Joke Reumers, Kyle L Morris, Alastair Copland, Louise Serpell, Luis Serrano, et al. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nature methods*, 7(3):237–242, 2010.
- [75] Scott McGinnis y Thomas L Madden. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic acids research*, 32(suppl 2):W20–W25, 2004.
- [76] Bálint Mészáros, Zsuzsanna Dosztányi, y István Simon. Disordered binding regions and linear motifs—bridging the gap between two models of molecular recognition. *PLoS ONE*, 7(10):e46829, 2012.
- [77] Bálint Mészáros, István Simon, y Zsuzsanna Dosztányi. Prediction of protein binding regions in disordered proteins. *PLoS computational biology*, 5(5):e1000376, 2009.
- [78] Amrita Mohan, Christopher J Oldfield, Predrag Radivojac, Vladimir Vacic, Marc S Cortese, A Keith Dunker, y Vladimir N Uversky. Analysis of molecular recognition features (MoRFs). *Journal of molecular biology*, 362(5):1043–1059, 2006.
- [79] Alexey G Murzin, Steven E Brenner, Tim Hubbard, y Cyrus Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, 247(4):536–540, 1995.
- [80] Hiroshi Nakashima y Ken Nishikawa. The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins. *FEBS letters*, 303(2):141–146, 1992.
- [81] Victor Neduvia, Rune Linding, Isabelle Su-Angrand, Alexander Stark, Federico De Masi, Toby J Gibson, Joe Lewis, Luis Serrano, y Robert B Russell. Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS biology*, 3(12):2090, 2005.

- [82] Rebecca Nelson, Michael R Sawaya, Melinda Balbirnie, Anders Ø Madsen, Christian Riekel, Robert Grothe, y David Eisenberg. Structure of the cross- β spine of amyloid-like fibrils. *Nature*, 435(7043):773–778, 2005.
- [83] Christopher J Oldfield, Yugong Cheng, Marc S Cortese, Pedro Romero, Vladimir N Uversky, y A Keith Dunker. Coupled folding and binding with α -helix-forming molecular recognition elements. *Biochemistry*, 44(37):12454–12470, 2005.
- [84] Christine A Orengo, AD Michie, S Jones, David T Jones, MB Swindells, y Janet M Thornton. CATH a hierachic classification of protein domain structures. *Structure*, 5(8):1093–1109, 1997.
- [85] Carlton Paul y Jürg P Rosenbusch. Folding patterns of porin and bacteriorhodopsin. *The EMBO journal*, 4(6):1593, 1985.
- [86] Pål Puntervoll, Rune Linding, Christine Gemünd, Sophie Chabanis-Davidson, Morten Mattingsdal, Scott Cameron, David MA Martin, Gabriele Ausiello, Barbara Brannetti, Anna Costantini, et al. ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic acids research*, 31(13):3625–3630, 2003.
- [87] Wei Qiao, Michelle Mooney, Amanda J Bird, Dennis R Winge, y David J Eide. Zinc binding to a regulatory zinc-sensing domain monitored in vivo by using FRET. *Proceedings of the National Academy of Sciences*, 103(23):8674–8679, 2006.
- [88] Vishnu Priyanka Reddy Chichili, Veerendra Kumar, y J Sivaraman. Linkers in the structural biology of protein–protein interactions. *Protein Science*, 22(2):153–167, 2013.
- [89] Rachel L Redler, David Shirvanyants, Onur Dagliyan, Feng Ding, Doo Nam Kim, Pradeep Kota, Elizabeth A Proctor, Srinivas Ramachandran, Arpit Tandon, y Nikolay V Dokholyan. Computational approaches to understanding protein aggregation in neurodegeneration. *Journal of molecular cell biology*, 6(2):104–115, 2014.
- [90] Clifford R Robinson y Robert T Sauer. Optimizing the stability of single-chain proteins by linker length and composition mutagenesis. *Proceedings of the National Academy of Sciences*, 95(11):5929–5934, 1998.

- [91] Christopher A Ross y Michelle A Poirier. Protein aggregation and neurodegenerative disease. 10(suppl):S10–S17, 2004.
- [92] Frederic Rousseau, Joost Schymkowitz, y Luis Serrano. Protein aggregation and amyloidosis: confusion of the kinds? *Current opinion in structural biology*, 16(1):118–126, 2006.
- [93] Michelle Sabourin, Creighton T Tuzon, Timothy S Fisher, y Virginia A Zakian. A flexible protein linker improves the function of epitope-tagged proteins in *Saccharomyces cerevisiae*. *Yeast*, 24(1):39–45, 2007.
- [94] Thomas D Schneider y R Michael Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic acids research*, 18(20):6097–6100, 1990.
- [95] Benjamin Schuler, Everett A Lipman, Peter J Steinbach, Michael Kumke, y William A Eaton. Polyproline and the “spectroscopic ruler” revisited with single-molecule fluorescence. *Proceedings of the National Academy of Sciences of the United States of America*, 102(8):2754–2759, 2005.
- [96] Joost Schymkowitz, Jesper Borg, Francois Stricher, Robby Nys, Frederic Rousseau, y Luis Serrano. The foldx web server: an online force field. *Nucleic acids research*, 33(suppl 2):W382–W388, 2005.
- [97] Christian JA Sigrist, Lorenzo Cerutti, Nicolas Hulo, Alexandre Gattiker, Laurent Falquet, Marco Pagni, Amos Bairoch, y Philipp Bucher. PROSITE: a documented database using patterns and profiles as motif descriptors. *Briefings in bioinformatics*, 3(3):265–274, 2002.
- [98] Jean D Sipe y Alan S Cohen. Review: history of the amyloid fibril. *Journal of structural biology*, 130(2):88–98, 2000.
- [99] Temple F Smith y Michael S Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.
- [100] Erik LL Sonnhammer, Gunnar Von Heijne, Anders Krogh, et al. A hidden markov model for predicting transmembrane helices in protein sequences. En *Ismb*, tomo 6, págs. 175–182. 1998.
- [101] Peter Tompa, Monika Fuxreiter, Christopher J Oldfield, Istvan Simon, A Keith Dunker, y Vladimir N Uversky. Close encounters of the third kind: disordered domains and the interactions of proteins. *Bioessays*, 31(3):328–335, 2009.

- [102] Ryan Trinh, Brian Gurbaxani, Sherie L Morrison, y Manouchehr Seyfzadeh. Optimization of codon pair use within the (GGGGS) 3 linker sequence results in enhanced protein expression. *Molecular immunology*, 40(10):717–722, 2004.
- [103] Antonio Trovato, Fabrizio Chiti, Amos Maritan, y Flavio Seno. Insight into the structure of amyloid fibrils from the analysis of globular proteins. *PLoS computational biology*, 2(12):e170, 2006.
- [104] Vladimir N Uversky, Joel R Gillespie, y Anthony L Fink. Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins: Structure, Function, and Bioinformatics*, 41(3):415–427, 2000.
- [105] Vladimir Vacic, Christopher J Oldfield, Amrita Mohan, Predrag Radivojac, Marc S Cortese, Vladimir N Uversky, y A Keith Dunker. Characterization of molecular recognition features, MoRFs, and their binding partners. *Journal of proteome research*, 6(6):2351–2366, 2007.
- [106] Robin van der Lee, Marija Buljan, Benjamin Lang, Robert J Weatheritt, Gary W Daughdrill, A Keith Dunker, Monika Fuxreiter, Julian Gough, Joerg Gsponer, David T Jones, et al. Classification of intrinsically disordered regions and proteins. *Chemical reviews*, 114(13):6589–6631, 2014.
- [107] Joost Van Durme, Sebastian Maurer-Stroh, Rodrigo Gallardo, Hannah Wilkinson, Frederic Rousseau, y Joost Schymkowitz. Accurate prediction of DnaK-peptide binding via homology modelling and experimental data. *PLoS computational biology*, 5(8):e1000475, 2009.
- [108] Kim Van Roey, Bora Uyar, Robert J Weatheritt, Holger Dinkel, Markus Seiler, Aidan Budd, Toby J Gibson, y Norman E Davey. Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chemical reviews*, 114(13):6733–6778, 2014.
- [109] Salvador Ventura, Jesús Zurdo, Saravanakumar Narayanan, Matilde Parreño, Ramón Mangues, Bernd Reif, Fabrizio Chiti, Elisa Giannoni, Christopher M Dobson, Francesc X Aviles, et al. Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case. *Proceedings of the National Academy of Sciences of the United States of America*, 101(19):7258–7263, 2004.
- [110] Jan L Vinkenborg, Tamara J Nicolson, Elisa A Bellomo, Melissa S Koay, Guy A Rutter, y Maarten Merkx. Genetically encoded FRET sensors to monitor intracellular Zn²⁺ homeostasis. *Nature Methods*, 6(10):737–740, 2009.

- [111] Christine Vogel, Matthew Bashton, Nicola D Kerrison, Cyrus Chothia, y Sarah A Teichmann. Structure, function and evolution of multidomain proteins. *Current opinion in structural biology*, 14(2):208–216, 2004.
- [112] Gunnar von Heijne. Membrane proteins: from sequence to structure. *Annual review of biophysics and biomolecular structure*, 23(1):167–192, 1994.
- [113] Ian Walsh, Flavio Seno, Silvio CE Tosatto, y Antonio Trovato. PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic acids research*, 42(W1):W301–W307, 2014.
- [114] Edward A Weathers, Michael E Paulaitis, Thomas B Woolf, y Jan H Hoh. Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. *FEBS letters*, 576(3):348–352, 2004.
- [115] Weiss, Manfred S. and Abele, Ulrich and Weckesser, Jürgen and Welte, Wolfram and Schiltz, Emile and Schulz, Georg E. Molecular architecture and electrostatic properties of a bacterial porin. *Science*, 254(5038):1627–1630, 1991.
- [116] Willy Wriggers, Sugoto Chakravarty, y Patricia A Jennings. Control of protein functional dynamics by peptide linkers. *Peptide Science*, 80(6):736–746, 2005.
- [117] Fan Xue, Zhong Gu, y Jin-an Feng. LINKER: a web server to generate peptide sequences with extended conformation. *Nucleic acids research*, 32(suppl 2):W562–W565, 2004.
- [118] Kai Yu, Chengcheng Liu, Byung-Gee Kim, y Dong-Yup Lee. Synthetic fusion protein design and applications. *Biotechnology advances*, 33(1):155–164, 2015.