Tesis de Doctorado

Programa de Desarrollo de Ciencias Básicas (PEDECIBA)

“Diseño e implementación de nuevas herramientas para la solubilización, evolución y cristalogénesis de proteínas”

Magister Agustín Correa

Tutor: Dr. Pedro M. Alzari

Co-Tutor: Dr. Pablo Oppezzo

Tribunal:

Dr. Alejandro Buschiazzo

Dr. Gualberto González

Dr. Pablo Aguilar

“Sin experimentación no hay verdad”

Aristóteles 384-322 a.C.

Contenido

RESUMEN:

El estudio de las proteínas y su rol biológico en la fisiología humana ha sido motivo de numerosas investigaciones tanto para las áreas de las ciencias básicas (estudios biofísicos, bioquímicos o estructurales) como aplicadas (biotecnología, salud humana, etc.). En este contexto la disponibilidad de la estructura tridimensional proteica, principalmente mediante la cristalografía de rayos X, ha sido fundamental para el entendimiento de las diversas funciones proteicas. Uno de los primeros impedimentos para el estudio funcional de las proteínas por técnicas cristalográficas ha sido la obtención de una proteína específica en estado puro y en cantidades suficientes. La posibilidad de alcanzar estas características raramente se logra directamente del huésped natural llevando a la producción de la proteína en forma recombinante como única alternativa posible. Dentro de los posibles huéspedes el más ampliamente utilizado es *Escherichia coli (E. coli)*, que si bien en muchos casos permite producir y purificar grandes cantidades de proteínas recombinantes (PRs), se estima que sólo un 30% de los genes clonados en esta bacteria se logran expresar de forma soluble y homogénea (1). Sin embargo, mediante la evaluación de diversas variables como ser diferentes promotores procariotas, proteínas de fusión, temperatura de inducción entre otras, es posible incrementar las probabilidades de obtener un producto soluble, homogéneo y en cantidades óptimas (2, 3). Parte de este trabajo de Tesis está abocado a generar metodologías que faciliten la obtención de proteínas solubles y en estado homogéneo.

Con el objetivo de evaluar la mayor cantidad de variables en el menor tiempo posible, es que en el presente trabajo se generó una serie de 12 vectores de expresión en *E. coli*, que permiten realizar un clonado con alta eficiencia en paralelo e independiente de la secuencia (4). La disponibilidad de esta técnica es un factor fundamental para cuando se trabaja con distintas construcciones y varias PRs en simultáneo. En esta tesis, además de generar esta serie de vectores, se propone el uso de una nueva proteína de fusión para ayudar en la expresión soluble de distintas PRs. Esta proteína de fusión corresponde a una versión truncada de la endoglucanasa D (CelDnc) de *Clostridium thermocellum*, siendo una proteína termostable y expresada en cantidades masivas en *E. coli* (5). Nuestro trabajo muestra que esta proteína de fusión puede llevar a una mejora en la obtención del producto soluble y que al menos para una de las proteínas estudiadas se comporta igual o mejor, que otras proteínas de fusión difundidas comercialmente como MBP, SUMO, DsbC y Trx.

Uno de los principales avances y aplicaciones más fascinantes de las PRs es su aplicación en la medicina. La generación de moléculas que pueden ser inyectadas en seres humanos con fines terapéuticos y/o de diagnóstico abre una puerta de gran interés en el campo de la medicina

(6). En este sentido, la generación de pequeñas moléculas (100-800 Da) que pueden funcionar como inhibidores de distintas enzimas y receptores ha tenido mucho éxito en el tratamiento contra diversos tipos de cáncer como es el caso de los inhibidores de tirosina quinasa: Erlotinib (Tarceva®), Imatinib (Gleevec®) y Dasatinib (Sprycel®) entre otros (7). Sin embargo este tipo de compuestos se encuentra con el problema que difícilmente pueden inhibir interacciones proteína-proteína, lo que muchas veces puede ser de gran utilidad en diversas terapias contra el cáncer (8). Los Anticuerpos monoclonales (AcMo), han surgido como una alternativa importante, cuando se requiere de interacciones más complejas o bien la activación del propio sistema inmune para la destrucción específica de la célula tumoral. Estas moléculas de mayor tamaño (>150 kDa) y mayor complejidad estructural y funcional están siendo cada vez más utilizadas en numerosas aplicaciones médicas para el tratamiento de diferentes patologías infecciosas, inmunes, cardiovasculares y diversos tipos de cáncer (9). Por ejemplo el anticuerpo quimérico Rituximab (Rituxan®), fue el primer anticuerpo aprobado para terapia contra el cáncer en 1997. Esta inmunoglobulina de isotipo IgG1, reconoce la proteína CD20 de la superficie de los linfocitos B, siendo utilizado con éxito en el tratamiento contra tumores de células B como el linfoma no-Hodgkin entre otros (9). Sin embargo, los AcMo no son de gran utilidad cuando se necesita inhibir la acción de una enzima cuyo sitio activo se encuentra en cavidades profundas, ya que éstos tienden a unir superficies planas (10). Esto abre una oportunidad para nuevas proteínas de unión de tamaño intermedio (6 a 20 kDa), capaces de interaccionar con regiones poco accesibles y a la vez por su mayor tamaño respecto a las pequeñas moléculas, conferir mayor especificidad de unión e interferir tanto en interacciones proteína-proteína como proteína-sustrato.

Es así que en esta tesis explotamos las características de proteínas de unión artificial como son las afitinas, derivadas de la proteína Sac7d (7 kDa), donde se utilizó un nuevo diseño de librería de manera de ampliar el repertorio de los posibles blancos a los que pueda unirse. En este sentido se utilizó además de la librería descripta previamente (11), una librería conteniendo un bucle o “loop” con mutaciones aleatorias, para poder interactuar con cavidades profundas en las proteínas como por ejemplo algunos sitios activos. Esto nos permitió obtener proteínas de unión termoestables con afinidades en el orden de nanomolar contra la glicosidasa CelD, que eran capaces además de inhibir su actividad tanto a 25˚ como a 60˚C. Al mismo tiempo también se realizaron ensayos de inhibición con otra afitina ya obtenida previamente capaz de unir a la glicosidasa lisozima y que provenía de una librería sin el mencionado loop (12). En este caso también se observó inhibición de la actividad de la lisozima. Las distintas afitinas no mostraron reactividad cruzada entre ambas enzimas a pesar de que las últimas presentan

sitios activos similares (13). Estudios estructurales de alta resolución con los diferentes complejos nos permitieron dilucidar el mecanismo de unión/inhibición, encontrándose que el modo de unión era diferente para las dos enzimas. De esta manera y utilizando diferentes formatos de librerías mostramos mediante ensayos bioquímicos y estructurales como un mismo esqueleto o "scaffold" puede unir e inhibir de manera potente y específica enzimas al unirse a superficies topológicamente diferentes. Por lo tanto, este grupo de librerías, es una buena herramienta para la obtención de proteínas de unión contra diversas superficies proteicas que podrán utilizarse en la inhibición de otros blancos biológicos y/o terapéuticos (manuscrito aceptado).

El uso de la cristalografía ha sido fundamental para entender diversos procesos biológicos, mecanismos de acción y el diseño basado en la estructura de distintas drogas de uso clínico (7). Sin embargo y a pesar de su importancia, además de contar con la proteína en estado soluble, es necesaria la formación de cristales para la obtención de la estructura cristalográfica. Esto último plantea una limitante ya que al día de hoy la cristalogénesis continúa siendo un proceso empírico de ensayo y error. Por ejemplo resultados de 7 centros de genómica estructural, mostraron que de 21149 genes clonados, el 36.2% se expresó en forma soluble y sólo se obtuvo cristales para el 7.3% (14). En este sentido es que nos propusimos utilizar a la proteína CelD, que mostró cristalizar de manera robusta (13, 15), para asistir en la cristalización de otras proteínas recombinantes. Como una primera prueba de concepto para esta hipótesis, mediante fusión directa con CelD, se logró resolver la estructura cristalográfica de un dominio catalítico (CA) de 17 kDa de la proteína de *E. coli* CpxA a 2.0 Å de resolución. Otro dato de interés que se desprende de estos resultados es que se logra obtener el cristal en la misma condición de cristalización que para CelD sola, es decir sin realizar un verdadero cribado de cristalización. Además, mediante ingeniería genética se optimizaron los contactos cristalinos para utilizar a CelD como fusión N-terminal en la cristalización de otras proteínas blanco. En otra estrategia, se generaron diversas librerías de mutantes del dominio tipo Ig de CelD para generar proteínas de unión contra una proteína blanco y de esta manera tras su unión cristalizar el complejo.

En el presente trabajo de Doctorado, se generó una herramienta que facilita la evaluación de condiciones de expresión de PRs, mostrándose además el uso de una nueva proteína de fusión capaz de aumentar la expresión y solubilidad de una proteína blanco. Por otro lado, se evaluó una nueva librería de mutantes de afitinas, que presenta un modo de unión que permite la interacción con cavidades profundas como las encontradas en los sitios activos de diversas

enzimas. Finalmente, también se evaluó la utilidad de la proteína CelD para su uso como proteína de fusión en la cristalización de otras proteínas blanco.

INTRODUCCIÓN

1. Expresión de proteínas recombinantes en *E. coli*, desafíos y soluciones

1.1 Generalidades

Con el surgimiento de la tecnología del ADN recombinante, en la década de 1970 se accede por primera vez a la capacidad de manipular genes a través de los procesos de clonado y expresión de ese gen a proteína. A partir de este momento la expresión de proteínas recombinantes (PRs) pasa a ser el método de elección en el área de la investigación científica y tecnológica debido a su reproducibilidad y su relativo bajo costo económico. La producción de PRs implica que el gen de interés puede ser introducido en hospederos (sistemas procariotas o eucariotas) donde la maquinaria de síntesis proteica de los mismos es redirigida hacia la expresión de la proteína blanco.

Esta tecnología ha tenido un impacto profundo en diversos campos de la ciencia, incluyendo el área de la medicina. La producción de PRs ha aportado conocimientos básicos y necesarios para el entendimiento de distintas enfermedades así como también han sido utilizadas directamente en la terapia y el diagnostico molecular. Por ejemplo, a comienzos de la década de 1980, la FDA (del inglés Food and Drug Administration) aprobó el uso clínico de la primera proteína recombinante producida en *Escherichia coli*, la insulina humana. Esta es utilizada para el tratamiento de la diabetes y su expresión de manera recombinante estableció los antecedentes de un modelo para el desarrollo de otros productos recombinantes terapéuticos (16).

Se han desarrollado diferentes huéspedes para la expresión de las PRs como ser células de mamífero, células de insecto y microorganismos. Dentro de ellos el más utilizado es la enterobacteria *E. coli*. Aproximadamente el 60% de las PRs reportadas en la literatura y el 30% de los productos recombinantes aprobados por la FDA son producidos en este huésped (17-19). El uso extendido de este sistema de expresión está principalmente asociado al gran rendimiento de producción, al fácil manejo e implementación en el laboratorio, a la disponibilidad de un gran número de herramientas genéticas y finalmente a su bajo costo. A pesar de estas características positivas, también existen desventajas asociadas al uso de *E. coli*.

Dentro de ellas debemos contemplar la falta de modificaciones post-traduccionales como la N- y O- glicosilación, amidación, entre otras (20) lo que dependiendo de la proteína blanco, puede limitar su uso. Además, la alta productividad de *E. coli* puede llevar a la acumulación del producto en el citoplasma de la bacteria en forma de agregados insolubles conocidos también como cuerpos de inclusión (21). Es así que se estima que sólo un tercio de los genes clonados en *E. coli* se logran expresar de forma soluble y homogénea (1).

La probabilidad de obtención de un producto puro y homogéneo, puede ser aumentada realizando una intensa optimización de las condiciones de expresión. Para ello diversas variables son evaluadas en combinación de manera de encontrar las condiciones de expresión óptimas para la proteína recombinante de interés. En este sentido, diferentes cepas de *E. coli* con diversas propiedades son evaluadas así como también distintas temperaturas de expresión, promotores, proteínas de fusión y chaperonas o interactores biológicos (2, 3) (ver anexo I y II). También la generación de diferentes variantes génicas (mutantes puntuales y versiones truncadas) de la proteína de interés pueden ser evaluadas. La combinación de todas estas variables genera un enorme número de condiciones a evaluar. En este sentido, los avances en robótica y miniaturización han sido esenciales para poder evaluar miles de condiciones de expresión diferentes, permitiendo un cribado de alto rendimiento (HTS, del inglés “High-throughput screening”) en pocos días y de manera automática (2).

Finalmente, además del proceso de evaluación de manera manual o por técnicas robóticas de numerosas variables de expresión existe otro recurso capaz de ofrecer una mejora en la producción de las PRs, la evolución dirigida. Esta consiste en la generación de una librería de mutantes de la proteína blanco y selección de aquellos que presentan el fenotipo deseado, en este caso la expresión soluble de la proteína de interés (22).

1.2 Primeros pasos a seguir para la expresión de una PR.

Una de las principales razones por la que muchas de las proteínas eucariotas no pueden expresarse de manera soluble en *E. coli*, es el requerimiento de modificaciones post-traduccionales para su correcto plegamiento. Una primera aproximación cuando se desea expresar la proteína de interés es el análisis de su secuencia nucleotídica y proteica. Existen servidores que contienen diferentes programas bioinformáticos para la predicción de la presencia de tales modificaciones (sitios de N- y O- glicosilación, fosforilación, etc), y la localización subcelular entre otros (http://www.expasy.org) (23). Esta información puede brindar una buena idea del éxito posible y ayudar en la estrategia a seguir para la producción de distintas PRs. Otros factores que pueden también tener un impacto y deben ser tenidos en

cuenta son: el uso de codones, la secuencia en la región de iniciación de la traducción (TIR, del inglés "Translation Initiation Region"), y la presencia de puentes disulfuro.

1.3 Promotores para la producción de PR y secuencias cercanas

El promotor a utilizar es de fundamental importancia, ya que controla la potencia y duración de la transcripción y por ende el rendimiento de producción. El promotor ideal debería tener una alta tasa de transcripción, presentar niveles muy bajos de expresión basal y ser inducido de una forma simple y eficiente. Los inductores pueden ser térmicos o químicos donde el análogo de la lactosa isopropil-β-D-tiogalactopiranósido (IPTG) es el inductor químico más ampliamente utilizado (24).

1.3.1 Promotores químicos

Dentro de los promotores químicos, el sistema de expresión comercializado bajo el nombre de pET (Novagen) basado en el promotor T7 es el más utilizado para la expresión de las PRs. Fue desarrollado por Studier y Moffatt (25, 26), y está basado en la alta selectividad de la ARN polimerasa del fago T7 para la transcripción del gen de interés. El promotor T7 es considerado un promotor fuerte, con una alta tasa de transcripción, alcanzando niveles de producción de la proteína recombinante de hasta un 50% del contenido total proteico. Debido a que *E. coli* carece de polimerasa para la transcripción de genes regulados por el promotor T7, es que algunas cepas como BL21(DE3), fueron modificadas para contener el gen de la RNA polimerasa T7 en el cromosoma bacteriano, bajo el control del promotor lacUV5 (25, 27). Dicho promotor tiene mutaciones puntuales que aumentan su eficiencia y lo hacen menos sensible a la represión por catabolito, por lo que se puede realizar la inducción con IPTG incluso en presencia de glucosa (28). De esta manera tras la adición del IPTG, la represión causada por la unión del LacI al operador lac presente en lacUV5 es liberada, resultando en la transcripción y traducción de la polimerasa T7 que a su vez transcribe el gen de interés con la concomitante producción de la proteína recombinante (29).

Otro promotor comúnmente utilizado es el promotor T5. El mismo puede ser reconocido por la RNA polimerasa de *E. coli* y posee un operador *lac* doble que provee de una regulación más precisa. Además es considerado un promotor menos fuerte que el T7 (30). El promotor T5 se encuentra en los vectores comerciales pQE (QIAGEN).

Finalmente el promotor arabinosa (P_{BAD}) también ha sido utilizado para la expresión de PRs. Cuando un gen es regulado por este promotor, la expresión es controlada por la proteína araC, la cual es un regulador positivo y negativo del promotor P_{BAD}. La inducción se logra con la adición de L-arabinosa (usualmente 0.2% w/v). Este es un promotor titulable, por lo que a

mayores concentraciones de inductor mayor expresión. Además es un promotor que está regulado de manera estringente y la expresión basal puede ser reducida aún más con la adición de glucosa, por lo que es ideal cuando se intenta expresar genes que pueden ser tóxicos para la célula (31, 32).

1.3.2 Promotores controlados por temperatura

En este tipo de promotores, la inducción se logra al cambiar la temperatura en lugar de necesitar la adición de un compuesto químico. A nivel industrial esto puede presentar una gran ventaja económica y operacional. Dentro de los promotores sensibles a la temperatura encontramos el promotor CspA (inducido a 15°C) y pL/pR del fago lambda (inducido a 40-42°C). CspA es la proteína más abundante en respuesta a bajas temperaturas en *E. coli*. La misma es prácticamente indetectable a 37°C y es intensamente producida al transferir los cultivos a 15°C (33). De esta manera los genes que se encuentran bajo el control de este promotor pueden ser inducidos al bajar la temperatura entre 15-25°C (34, 35). Los promotores pL/pR del fago lambda, son regulados por una versión mutante sensible a la temperatura del represor cI857 del bacteriófago λ (36). En este caso, a bajas temperaturas (28-32°C), la transcripción es inhibida por la unión de cI857 al promotor pL o pR. Luego de aumentar la temperatura a 40-42°C, la unión del represor es liberada, dando lugar a la expresión del gen de interés (36-38).

1.3.3 Otras secuencias nucleotídicas que pueden afectar la expresión de la PR

Finalmente, también a nivel nucleotídico, la secuencia hacia el extremo 5' del gen puede tener un efecto importante en los niveles de expresión proteica debido a la generación de estructuras secundarias en el ARN mensajero (ARNm). Esto juega un papel importante en detrimento de la traducción por el complejo ribosomal. En este sentido, se mostró que secuencias ubicadas inmediatamente después del codón de iniciación hasta la posición +25 pueden afectar los niveles de expresión. Es así que algunos programas permiten definir mutaciones silenciosas en los primeros 7 codones para mejorar los niveles de expresión (39). Más recientemente se desarrolló un método predictivo para el diseño de sitios sintéticos de unión al ribosoma (RBS), que presentan diferentes tasas de traducción, permitiendo un ajuste fino de los niveles de expresión (https://salis.psu.edu/software) (40, 41). Esto se logra al optimizar varios parámetros como la disminución de estructuras secundarias en el ARNm transcrito, la distancia del RBS con el codón de iniciación, las interacciones moleculares entre el ARNm y el complejo ribosomal 30S entre otras (40, 41). En el mismo sentido, la vida media del ARNm en bacterias es más corta que en eucariotas. Esto se logra mejorar mediante una

mutación en el gen que codifica para la RNasaE confiriendo mayor estabilidad al ARNm (42). Dicha mutación se encuentra en la cepa BL21 Star (Invitrogen).

1.4 Cepas de *E. coli* utilizadas para la expresión de PRs

Diversas cepas de *E. coli* han sido desarrolladas y son comercializadas para la expresión de PRs. La cepa a utilizar tiene un profundo impacto en la producción de la proteína blanco ya que es la que aporta el contexto genético en el que ocurre la expresión. La selección de la cepa depende de las características de la proteína recombinante, como ser si la misma es tóxica para la bacteria, si presenta codones poco frecuentes debido a su origen heterólogo o si contiene puentes disulfuro, entre otras.

1.4.1 Cepas deficientes en proteasas y control de la expresión basal de la PR

Una de las cepas más comúnmente utilizadas en la producción de las PRs es BL21(DE3). Esta cepa es deficiente en las proteasas Lon y OmpT, lo cual reduce la degradación proteolítica de las PRs e incrementa el rendimiento final (43, 44). Además contiene una copia cromosómica del gen de la RNA polimerasa T7 bajo el control del promotor lacUV5, haciéndola adecuada para la expresión de genes controlados por el promotor T7. El alto rendimiento de la polimerasa T7 puede llevar a efectos tóxicos para la bacteria, reduciendo sustancialmente los rendimientos finales de las PRs. Con el objetivo de sortear este problema se originan las cepas BL21(DE3)pLysS y BL21(DE3)pLysE (Novagen) que se caracterizan por poseer un plásmido que codifica para la lisozima T7, inhibidor natural de la polimerasa T7. Estas variantes de BL21 son utilizadas generalmente cuando se trabaja con genes tóxicos por lo que la presencia de la lisozima T7 disminuye la expresión basal de la proteína recombinante y por ende sus efectos negativos para la bacteria (45). Mientras pLysS produce bajos niveles de lisozima T7, pLysE provee de mayores niveles del inhibidor. Otra aproximación empleada para la expresión de genes tóxicos, ha sido la de sustituir el promotor lacUV5 por un promotor mas astringente como el P_{BAD}, para el control de la polimerasa T7. Este es el caso de la cepa BL21AI (Invitrogen), donde la inducción debe realizarse con la adición de 0,2% L-arabinosa. Es importante notar en este último caso, que si se trabaja con un vector que expresa el represor LacI, es necesario agregar también IPTG para la inducción del gen de interés (31).

1.4.2 Expresión de genes con codones pocos comunes en E. coli

Debido a la naturaleza heteróloga de la proteína blanco, el gen de interés puede contener codones que se encuentran en baja abundancia en el hospedero. Esta cantidad relativa de codones y la poca disponibilidad de los ARN de transferencia específicos para los mismos lleva a la terminación prematura de la traducción y a bajos rendimientos en la producción de la proteína recombinante (46).

Existen dos estrategias principales para solucionar este problema: i) mediante la optimización de la secuencia del gen en forma racional; ii) mediante el uso de cepas de *E. coli* suplementadas con los tRNAs que se encuentran en baja abundancia. La optimización racional consiste en la substitución de los codones raros del gen de interés por codones utilizados con alta frecuencia en *E. coli*, mediante la síntesis *de novo* del gen. Varios algoritmos han sido desarrollados para optimizar la secuencia del gen al uso de codones del hospedero en donde se va a expresar la proteína recombinante (47, 48). Esto ha permitido la expresión en forma soluble de proteínas que de otra forma se expresaban como cuerpos de inclusión (49). Para la segunda estrategia, se encuentran varias cepas comercialmente disponibles que co-expresan tRNAs para los codones pocos frecuentes como ser BL21 CodonPlus (Novagen), Rosetta (Invitrogen) y BL21(DE3)RIL (Stratagene) entre otras. El uso de este tipo de cepas ha mejorado la expresión de varios genes humanos (50, 51).

Finalmente, si bien el cambio de los codones raros puede incrementar la tasa de traducción, en algunos casos podría llevar a la agregación y mal plegado de la proteína recombinante. Esto fue mostrado para varias proteínas eucariotas expresadas en *E. coli* donde la expresión soluble en BL21(DE)pLysS fue ampliamente mayor a la obtenida en BL21(DE3)CodonPlus-pRIL (52), sugiriendo que las pausas en la traducción son necesarias en algunos casos para el correcto plegado de dominios individuales de la proteína blanco (53).

1.4.3 Expresión de proteínas con puentes disulfuro

Los puentes disulfuro (S-S) se forman por enlace covalente entre dos átomos de azufre provenientes de 2 residuos de cisteína. Éstos son generalmente esenciales para el correcto plegamiento y estabilidad de la proteína por lo que es un elemento muy importante a tener en cuenta cuando se va a realizar la expresión de PRs (54). Actualmente existen varios programas disponibles en la web que permiten estimar con cierto grado de confianza la presencia de S-S en la proteína de interés (55, 56).

La formación de enlaces S-S requiere de un ambiente oxidante como el que se encuentra en el retículo endoplásmico de los eucariotas o en el periplasma bacteriano. En el caso de *E. coli*, existe una maquinaria específica para la formación de los puentes disulfuro conocida como sistema Dsb (del inglés “Disulfide bond”). Éste incluye las enzimas periplásmicas DsbA (cataliza la formación de los S-S) y DsbC (realiza la isomerización de los S-S). Luego el ciclo es reiniciado por la acción de las proteínas de membrana DsbB y DsbD que reciclan a DsbA y DsbC respectivamente (57).

Las proteínas son dirigidas al periplasma bacteriano para la formación de S-S al adicionar un péptido líder en el extremo amino terminal, que puede ser removido por una endopeptidasa una vez en el periplasma. Dentro de los péptidos lideres utilizados para la exportación periplásmica encontramos los de las proteínas DsbA, PelB, LamB, PhoA y OmpA entre otros (58, 59). Existen distintos mecanismos de exportación periplásmica como la vía TAT (del inglés "Twin-Arginine Translocon) que transloca proteínas plegadas (60) y el translocón Sec donde las proteínas son exportadas en forma desplegada. Además el translocón Sec está formado por la vía post-traduccional dependiente de SecB y la vía co-traduccional dependiente de SRP (del inglés "Signal Recognition Particle") (61). Una de las ventajas adicionales de la expresión periplásmica es el hecho de que la purificación de la PR es más fácil debido a la menor cantidad de proteínas de *E. coli* en este compartimento (62). Sin embargo, la maquinaria de tanslocación puede saturarse, lo cual es tóxico para la bacteria disminuyendo en gran medida el rendimiento final de la producción de la proteína recombinante (63). Esto puede ser aliviado (como se verá más adelante), mediante el uso de la cepa Lemo21(DE3) (59, 63).

Como alternativa a la exportación al periplasma para la formación de puentes disulfuro, diferentes cepas que presentan un citoplasma más oxidante han sido diseñadas, por ejemplo Origami 2(DE3) y Origami B(DE3) (Novagen). Estas cepas contienen mutaciones en los genes glutatión reductasa (*gor*$^-$) y tiorredoxina reductasa (*trxB*$^-$) que están involucradas en el mantenimiento de un ambiente reducido en el citoplasma así como también una mutación en el gen de la peroxirredoxina (*ahpC*) esencial para el crecimiento de los mutantes (54, 58). Una desventaja es que debido a las mutaciones en *gor* y *trxB*, las células muestran un crecimiento más lento lo que a su vez se refleja en un menor rendimiento de las PRs. Si bien estas mutaciones permitirían la formación de enlaces disulfuro en el citoplasma de *E. coli*, las proteínas con varios S-S necesitan además de una correcta isomerización de los mismos. Para solucionar este problema, se desarrolló la cepa Shuffle (New England Biolabs), en donde la isomerasa DsbC es expresada de forma constitutiva en el citoplasma de *E. coli,* además de contener las mutaciones *gor* y *trxB* (64).

Recientemente se logró producir PRs con puentes disulfuro correctamente formados, en el citoplasma de *E. coli* sin la necesidad de las mutaciones *gor*$^-$ y *trxB*$^-$ mencionadas. Esto fue logrado tras la co-expresión en otro plásmido de las proteínas sulfidril oxidasa de *Saccharomyces cerevisiae,* Erv1p (catalizador de la formación de puentes disulfuros) y la isomerasa disulfuro DsbC de *E. coli* (65, 66). Además de esta observación se encontró que, en algunos casos, la adición de Erv1p puede ser más efectiva que la utilización de las cepas Origami o Shuffle (65, 66). En un trabajo reciente, luego de hacer fusiones amino terminales

con DsbC y 28 proteínas pequeñas ricas en S-S, se encontró que la cepa BL21(DE3)pLysS fue mucho más eficiente que las cepas Origami B (DE3)pLysS o Shuffle T7Express lysY en producir formas oxidadas y correctamente plegadas de las proteínas. Interesantemente, se encontró que dichos puentes disulfuro se formaban *ex vivo*, durante los pasos de extracción y purificación de las PRs (67).

En conclusión, la expresión de proteínas con S-S presenta a veces un problema complejo que debe ser tenido en cuenta a la hora de producir una proteína recombinante. Sin embargo, varias alternativas pueden ser contempladas: **a)** exportación al periplasma, **b)** expresión en el citosol de una bacteria con las vías reductoras inactivadas, **c)** co-expresión de proteínas para la oxidación e isomerización de los S-S o **d)** fusión covalente con la isomerasa DsbC..

1.4.4 Expresión de proteínas de membrana

Se calcula que un 30% de los genes humanos codifican para proteínas integrales de membrana (IMP, del inglés "Integral Membrane Proteins") (68) y están involucradas en procesos celulares críticos incluyendo la comunicación celular, transporte de moléculas a través de la bicapa lipídica, adhesión celular, entre otras. Las proteínas de membrana constituyen una clase principal de blancos terapéuticos representando el 80% de los fármacos más vendidos en USA en 2010 (69). A pesar de la importancia de su estudio, han sido muy difíciles de caracterizar debido a su carácter hidrofóbico, que las hace sumamente inestables en solución, dificultando su purificación y cristalización. Esto último queda reflejado en la observación de que menos del 1% de las proteínas depositadas en la PDB (Protein Data Bank) corresponden a IMPs. *E. coli* es el huésped más utilizado para el caso de las proteínas de membrana de origen tanto procariota como eucariota (70). Usualmente luego de ser producidas y transportadas a la membrana de la bacteria, las IMPs deben ser retiradas de la bicapa lipídica mediante el uso de detergentes para sus posteriores análisis bioquímicos y estructurales. Las IMPs solubilizadas en detergentes quedan envueltas en micelas de detergente lo que mimetiza en parte la bicapa lipídica (69). Existen muchos tipos de detergentes que pueden ser utilizados para la solubilización de proteínas de membrana. Además de ello con frecuencia diferentes construcciones génicas o mutantes de la proteína nativa deben ser evaluadas para encontrar alguna variante que presente buena solubilidad y estado homogéneo, característica fundamental para el proceso de cristalogénesis. Por lo tanto, un número sumamente elevado de variables deben ser evaluadas para determinar las condiciones óptimas en la cual la proteína recombinante se mantenga en solución y pueda utilizarse en los ensayos de cristalogénesis. Una interesante aproximación para analizar múltiples variables en forma efectiva, ha sido descrita por Kawate *et al.* (71). En esta aproximación, se realiza una fusión directa de la IMP con GFP (del inglés

"Green Fluorescent Protein") y se acopla la detección de fluorescencia con cromatografía de exclusión molecular (SEC). De esta manera, es posible trabajar con extractos crudos, sin purificar y cantidades mínimas de proteína (niveles de nanogramos) para la determinación de niveles de expresión, grado de monodispersión y peso molecular aproximado (71). Recientemente, una aproximación mejorada de este sistema fue desarrollada. En este caso en lugar de fusionar la proteína a GFP, se utiliza una sonda que reconoce específicamente el His-Tag y fluorese de similar manera que GFP. De esta manera se evita la fusión con GFP, y por ende algunos problemas que pueden estar asociados por dicha fusión como la presencia de falsos positivos, o agregación de la IMP luego del corte proteolítico de la fusión (72).

Finalmente, existen cepas de *E. coli* que han mostrado ser más apropiadas para la expresión de IMP's, como son los mutantes derivados de BL21(DE3) conocidas como C41(DE3) y C43(DE3) (Lucigen) (73). Al secuenciar el genoma de dichas cepas se encontró que la razón de la mejora en la expresión de proteínas de membrana es el resultado de mutaciones en el promotor lacUV5 que disminuyen la expresión de la polimerasa T7. De esta manera se disminuye también la tasa de expresión de la proteína blanco, evitándose la saturación de la maquinaria de translocación de las PRs a la membrana (63). Esto llevó al desarrollo de una nueva cepa derivada de BL21(DE3), llamada Lemo21(DE3) (New England Biolabs). En esta cepa, existe un plásmido que codifica para la lisozima T7 bajo el control del promotor titulable de L-ramnosa. De esta forma, a mayores niveles de ramnosa, se produce una mayor cantidad de lisozima T7, la que al inhibir la polimerasa T7 produce una menor transcripción y expresión del gen de interés (63). Mediante el uso de esta cepa se pudo controlar de forma precisa los niveles de expresión de 2 proteínas blanco y evitar la saturación de la maquinaria de translocación (59).

Es importante notar también que si se trabaja con una IMP eucariota, hay que considerar las modificaciones post-traduccionales que en muchos casos pueden ser necesarias para su estabilidad y el correcto plegamiento (74). Además, la tasa de elongación de la cadena polipetidica es entre 4 a 10 veces más rápida en procariotas comparado con eucariotas. Esto puede llevar a la exposición de regiones hidrofóbicas y agregación de la proteína por lo que cuando se trabaja con IMPs es recomendable utilizar promotores de baja tasa de transcripción, en plásmidos con bajo número de copias y trabajar a bajas temperaturas de inducción (70, 75).

1.5 Condiciones de cultivo

Una estrategia común para la expresión de una PR en forma soluble es la de evaluar diferentes condiciones de cultivo como temperatura y tiempo de inducción y la composición del medio de cultivo.

Se ha encontrado frecuentemente que bajar la temperatura de inducción (16-25˚C), puede aumentar el rendimiento final de la proteína recombinante, debido a como se mencionó con anterioridad, una disminución en las tasas de traducción (76). Sin embargo, bajar la temperatura también puede disminuir la biomasa final, por lo que si la proteína se expresa con buenos rendimientos, es conveniente realizar la expresión a 37˚C.

Varios medios de diferente composición han sido utilizados y comparados para la expresión de PRs: LB (Luria Bertani), 2YT, TB (Terrific Broth), SB (Super Broth) y autoinducción. Dentro de éstos, el medio de autoinducción desarrollado por Studier (77), ha sido utilizado con gran éxito en la producción de PRs en un amplio rango de escalas. Este medio permite la obtención de altas densidades bacterianas sin la necesidad de monitorear el crecimiento celular. Además, debido a la presencia de glucosa en su composición, presenta un control más ajustado de la expresión basal. Estas características permitieron su fácil adaptación a placas de 96 y 24 pocillos para experimentos de tipo HTS (77-79). Finalmente existe una versión modificada del medio de autoinducción adaptada para sistemas de expresión basados en el promotor P_{BAD} (77).

Por otro lado, la adición de diferentes aditivos al medio de cultivo puede también tener un efecto en la solubilidad de las PRs, como fue mostrado recientemente con proteínas expresadas en el periplasma de *E. coli* (80). En este trabajo se realizó la expresión de un mutante inestable, Im7-I22V, como parte de una fusión tripartita con el gen reportero de la β-lactamasa. De esta manera si la proteína permanece inestable, provoca una disminución en la actividad del reportero, mientras si se logra estabilizar, la actividad se incrementa. Luego de evaluar varios osmolitos en el medio de cultivo, encontraron que la actividad pudo incrementarse de forma dosis-dependiente (hasta 207 veces al adicionar glicerol 2M) mostrando su potencial utilidad en las formulaciones de los medios de cultivo (80).

1.6 Proteínas de Fusión o “tags”.

Las proteínas de fusión (y también péptidos o “tags”), han sido ampliamente utilizadas para resolver dos obstáculos mayores en el campo de las PRs como son, aumentar la solubilidad y facilitar la purificación de la proteína blanco. Estas fusiones y/o “tags”, pueden ser separados en 3 categorías principales: i) Tags de afinidad, ii) proteínas potenciadoras de la solubilidad y iii) Fusiones de doble propósito (purificación y mayor solubilidad). Los tags de afinidad, son generalmente cortos y pueden localizarse tanto en el extremo amino terminal como carboxilo terminal. Estos son reconocidos por diferentes matrices de cromatografía permitiendo la purificación de la proteína blanco. Las proteínas potenciadoras de la solubilidad son proteínas

extremadamente solubles, que pueden ser termoestables presentar actividad chaperona y son usualmente utilizadas como fusiones N-terminal mejorando el plegamiento y solubilidad de la proteína blanco (2, 3).

1.6.1 Tags de afinidad

Dentro de los tags de afinidad el más utilizado es el HisTag, formado usualmente por 6 residuos de histidina en tándem (0.84 kDa). Este tag interacciona de manera reversible con iones metálicos como ser Ni o Co (y otros metales de transición), que se encuentran inmovilizados en una matriz capaz de coordinarlos (resinas Sepharose 6, GE; o resinas Talon, Clontech)(81). Las purificaciones a través del HisTag se conocen como IMAC (del inglés, Immobilized Metal ion Affinity Chromatography). Luego de la interacción, las proteínas unidas pueden ser eluídas utilizando un competidor como el imidazol, bajando el pH o utilizando un agente quelante como el EDTA. Una de las ventajas del HisTag es que no es necesaria una estructura terciaria determinada para la coordinación del metal, posibilitando la purificación incluso en condiciones desnaturalizantes (82, 83). Como desventaja, se ha encontrado que para aquellas proteínas de baja expresión fusionadas a un HisTag, el aumento del volumen de cultivo no siempre se correlaciona con un aumento en la recuperación de producto. Esto se debe a que el aumento de la biomasa lleva a un incremento de agentes quelantes presentes en el periplasma de *E. coli* que disminuyen la capacidad de unión de las resinas de purificación (84). Este efecto puede reducirse al remover el material periplásmico antes de la lisis celular. Además de esto se debe tener en cuenta la presencia de varias proteínas nativas de *E. coli* (como ArnA, SlyD y GlmS) que pueden unirse a las resinas de purificación por IMAC, especialmente cuando se trabaja con proteínas blanco que se expresan en bajos rendimientos (85). En este sentido, se han desarrollado cepas de *E. coli* que son mutantes en estas proteínas (NiCo21, New England Biolabs), disminuyendo los contaminantes que co-purifican con la proteína de interés (86, 87).

Otro tag de afinidad muy utilizado es el StrepTag II, que está formado por 8 residuos (WSHPQFEK) y es reconocido de manera reversible por una versión modificada de la estreptavidina (Strep-Tactin) (88). La elución se realiza con un competidor como la biotina o preferentemente la destiobiotina. Si bien la capacidad de unión por mililitro de resina es menor comparado con resinas para IMAC, presenta mayor especificidad por lo que es una muy buena opción a tener en cuenta cuando se trabaja con proteínas que se expresan en bajos rendimientos (88, 89).

1.6.2 Proteínas potenciadoras de la solubilidad

La fusión de PRs con proteínas potenciadoras de la solubilidad, ha permitido en muchos casos la expresión soluble de proteínas que de otra forma se expresaban insolubles. Además, en muchos casos, tras separar a la proteína blanco de la fusión mediante corte proteolítico, la misma permanecía en estado soluble y homogéneo, demostrando el potencial de esta aproximación (67, 78, 90-92). Dentro de las proteínas de fusión más utilizadas para la expresión soluble de PRs encontramos la proteína MBP (del inglés, maltose-binding protein), GST (glutatión-S-transferasa), Trx (tioredoxina A), DsbC (disulfuro isomerasa C), SUMO (del inglés, small ubiquitin-like modifier protein) y NusA (del inglés, N-utilization substance A). Además, las proteínas MBP y GST pueden también ser utilizadas como tags de afinidad funcionando así como fusiones doble propósito. MBP es una proteína periplásmica de *E. coli* de 42 kDa que se une fuertemente a resinas de amilosa, eluyendo por competición con maltosa libre (93). MBP es una de las proteínas de fusión más populares siendo utilizada como fusión tanto N- como C- terminal, derivando en el correcto plegamiento y aumento de solubilidad de muchas proteínas eucariotas (94-96). Además, si el péptido líder natural de MBP se mantiene en la construcción, la fusión es dirigida al periplasma de *E. coli*. La proteína GST de *Schistosoma japonicum* de 26 kDa, puede unir resinas con glutatión, y la elusión se logra tras la aplicación de glutatión reducido (97). A pesar de que GST es ampliamente utilizada, ha demostrado ser un solubilizador pobre, donde frecuentemente luego del corte proteolítico, la proteína blanco tiende a precipitar (78, 90, 94). Sin embargo en algunos casos permitió la solubilización de la proteína blanco y el método de purificación por GST todavía aparece como una opción atractiva por su alta especificidad.

Otra proteína de fusión es la oxidoreductasa Trx de *E. coli* de 11.6 kDa, una proteína termoestable (T_m: 85˚C) y extremadamente soluble cuando es expresada en *E. coli*, alcanzando hasta un 40% de las proteínas totales (98). La fusión con Trx ha mejorado el plegamiento y estabilidad de algunas proteínas blanco (99-101). También la isomerasa de puentes disulfuro DsbC de *E. coli* de 25 kDa, ha sido utilizada con éxito para el correcto plegamiento de proteínas con varios enlaces S-S como se mencionó con anterioridad (58, 67, 102). Esta proteína muestra actividades de isomerasa de S-S y chaperona. SUMO es otra proteína de fusión de 11.2 kDa de origen eucariota (levadura), cuando es utilizada como fusión N-terminal durante la expresión procariota puede promover el correcto plegado y aumentar la solubilidad de la proteína blanco (103-105). Además las fusiones con SUMO pueden ser cortadas por la proteasa específica Ulp1, que reconoce elementos de la estructura terciaria y un motivo Gly-Gly en el extremo C-

terminal de SUMO, dejando un amino terminal nativo en la proteína blanco, excepto por prolina (104).

Finalmente, el factor de elongación de transcripción y antiterminación de *E. coli*, NusA (55 kDa), también ha mostrado ser un buen potenciador de la solubilidad para varias PRs (106). En un estudio comparativo usando varios blancos proclives a la agregación, las propiedades solubilizadoras de NusA fueron comparables a las de MBP (107).

Debido a que las proteínas de fusión Trx, DsbC, Sumo y NusA no facilitan la purificación, son generalmente utilizadas en combinación con tags de afinidad como HisTag o StrepTag II. Por otro lado, no existe al día de hoy una fusión que sirva para todas las PRs, por lo que varias fusiones diferentes deben evaluarse para aumentar las probabilidades de éxito en la expresión soluble de una PR.

Por último, estas proteínas de fusión o potenciadores de solubilidad, no sólo han sido útiles para propósitos de expresión/purificación, sino también para la obtención de estructuras cristalográficas cuando estaban fusionadas a proteínas blanco. Entre las proteínas de fusión que han sido utilizadas para cristalizar proteínas blanco encontramos a MBP, GST, Trx y GFP (108-111).

1.6.3 Corte proteolítico de la fusión o tag

Todos los tags de fusión pueden interferir con posteriores estudios estructurales y funcionales de la proteína de interés. Generalmente, luego de la purificación de la fusión completa, el tag es eliminado con una endoproteasa que corta en una secuencia específica situada entre el tag y la proteína blanco.

Existen varias endoproteasas disponibles para el corte proteolítico como ser enteroquinasa, factor Xa y trombina, pero la más utilizada es la proteasa del virus del tabaco (TEV) debido a que posee varias ventajas. Entre ellas podemos destacar su especificidad (reconoce el sitio ENLYFQ'G), el corte puede realizarse a 4°C y es producida con altos rendimientos en *E. coli* (112). Además, si bien su máxima actividad se da en condiciones reductoras (típicamente 1 mM DTT), el corte todavía ocurre en ausencia de reductor, lo que puede ser importante si se trabaja con proteínas con S-S (113).

1.7 Evaluación de la expresión en forma HTS

Cuando se trabaja con una proteína difícil de expresar, la evaluación de las variables mencionadas anteriormente (temperatura, promotor, proteínas de fusión, etc) puede ser esencial para la obtención de la proteína de interés en forma soluble. Esto lleva a que deban

evaluarse cientos de condiciones diferentes. Actualmente es posible lograr dicha evaluación en tiempo y costos razonables mediante el cribado de alto rendimiento o HTS. La generación de estaciones robóticas para el pipeteo ("liquid handling workstations"), permitieron trasladar los avances de automatización al laboratorio. Las metodologías de HTS combinan la automatización con la miniaturización, permitiendo la evaluación automática de miles de condiciones, en una escala de tiempo de semanas o incluso días. La implementación de este tipo de tecnologías ha permitido encontrar la combinación óptima de vector, cepa, condición de cultivo o condición de renaturalización requerida para el correcto plegamiento de diferentes PRs (114-117). Con el desarrollo de pipetas multicanal más sofisticadas, también es posible evaluar muchas condiciones pero de manera manual, pudiendo implementarse en laboratorios con menor equipamiento.

En este tipo de protocolos, las bacterias de *E. coli* son cultivadas en placas de 24 pocillos conteniendo 4 ml de medio como ser autoinducción o TB para una máxima obtención de biomasa (2). Luego de la inducción, la lisis celular puede realizarse con tampones comerciales como ser PopCulture™ (Novagen) o FastBreak™ (Promega) (118, 119) o mediante la adición de lisozima, magnesio y DNAsaI combinado con un ciclo de congelado descongelado (5). Los sobrenadantes obtenidos por centrifugación o los extractos totales son purificados directamente en placas de 96 pocillos por ejemplo conteniendo bolillas cargadas con níquel para realizar una purificación por IMAC (81, 115, 120). Luego de la elución de la proteína con imidazol y mediante centrifugación o aplicación de vacío, se puede evaluar la expresión de las diferentes condiciones utilizando geles de 96 pocillos como ser E-PAGE™96 system (Invitrogen) (5, 118, 119). De esta manera y utilizando materiales relativamente simples es posible tener la capacidad de evaluar un número importante de condiciones de expresión.

1.8 Renaturalización de cuerpos de inclusión

Frecuentemente las PRs se acumulan como agregados insolubles en el citoplasma de *E. coli* conocidos como cuerpos de inclusión (CI). Si bien puede parecer muy inconveniente, la formación de CI presenta algunas ventajas. Cuando las proteínas se expresan como CI, lo hacen con altos rendimientos y alta homogeneidad en composición donde en algunos casos la proteína recombinante puede representar más del 90% de las proteínas totales en esta fracción, facilitando la purificación de la proteína luego de su solubilización (121). Las condiciones para la renaturalización de la proteína recombinante, incluyen la evaluación de varios parámetros incluyendo pH, fuerza iónica, temperatura, adición de compuestos de bajo peso molecular, entre otros. En este sentido, varias aproximaciones se han utilizado que incluyen formatos en placas de 96 pocillos, para facilitar la optimización de las condiciones de

plegado de manera “HTS” (116, 122).También existen diferentes métodos para el proceso de renaturalización como ser dilución, diálisis y plegado en columna lo que puede tener efectos diversos en el plegamiento de la proteína recombinante (82, 116, 123).

Si bien lo habitual es intentar expresar la proteína en forma soluble, existe otra estrategia en donde se fusiona un péptido para reducir la solubilidad de la proteína blanco y dirigirla a CI. Esto puede ser útil por ejemplo cuando la proteína de interés es tóxica para la bacteria en su estado soluble. Un ejemplo concreto de proteína de fusión para este propósito, es el de una versión mutante de la autoproteasa del virus de la fiebre clásica porcina Npro, llamada EDDIE. Cuando ésta se encuentra fusionada al extremo amino terminal de la proteína recombinante, puede reducir su solubilidad favoreciendo la formación de CI. Al cambiar de condiciones caotrópicas a cosmotrópicas, la proteasa se vuelve activa cortando la fusión y liberando de esta manera la proteína recombinante con el extremo amino terminal nativo (124, 125).

Por último, la percepción de la naturaleza de los CI, ha cambiado dramáticamente en los últimos años. Usualmente se asumía que los CI estaban formados por agregados inertes compuestos por proteínas parcial o totalmente desplegadas, en lugar de moléculas con un plegado nativo y maduro. Sin embargo, en las últimas décadas, se concluyó en varios casos que los CI podían estar constituidos por proteínas activas (126-129). Esto abrió la posibilidad de utilizar los CI directamente, sin realizar la renaturalización, en aquellas aplicaciones donde la agregación de la proteína no es un impedimento. De esta manera se facilita en gran medida la producción/purificación, reduciendo enormemente los costos (130-132).

1.9 Evolución dirigida para la expresión soluble de PR

A pesar de la evaluación de gran cantidad de condiciones de expresión para las PRs, muchas veces no es posible encontrar la condición en donde la proteína blanco se exprese de manera soluble y homogénea. En estas situaciones, en lugar de explorar más condiciones de expresión, puede ser más eficiente cambiar las características bioquímicas de la proteína, mediante mutaciones y/o deleciones en la secuencia nucleotídica, para mejorar su solubilidad/estabilidad. Cuando se dispone de información estructural (por ejemplo de una proteína muy relacionada o cercana evolutivamente), estas modificaciones pueden lograrse de forma racional mediante mutagénesis dirigida (133). Desafortunadamente, en la mayoría de los blancos interesantes, esta información estructural no está disponible, viéndose limitado el uso de un diseño racional. En este escenario, una alternativa posible es la de realizar evolución dirigida. Esta aproximación está basada en un proceso iterativo que consiste en un primer paso de diversificación seguido de un paso de selección de los mutantes que presentan mejoras en

el fenotipo deseado. El paso de diversificación, se logra usualmente mediante mutaciones aleatorias (PCR proclive al error, mutagénesis química, uso de cepas de *E. coli* con altas tasas de mutación) y/o recombinación *in vitro* ("DNA shuffling") (134), generándose una librería de mutantes. Luego del paso de diversificación, se deben seleccionar esos pocos mutantes que presentan en este caso una mejora en la solubilidad/estabilidad, entre los millones de mutantes fútiles. Esta selección se logra al analizar la actividad de una proteína reportera, que está asociada a la solubilidad de la proteína blanco, o si es posible la actividad de la proteína blanco (135).

Un reportero del plegamiento que ha sido utilizado con éxito para lograr la evolución de variantes activas y solubles es la proteína GFP (136, 137). En este sistema la proteína blanco es expresada como fusión N-terminal con GFP, acoplando de esta forma la fluorescencia de las células de *E. coli* con el plegamiento productivo de la proteína fusionada (136). El aislamiento de las células más brillantes en busca de las mutaciones en el blanco que lo hacen más soluble, se puede realizar por análisis de colonias o separación de células por fluorescencia activada en un citómetro de flujo (FACS, del inglés, fluorescence-activated cell sorting). Este sistema se logró optimizar, luego del diseño de una forma partida de GFP (Split-GFP), auto complementaria (138), derivada de una variante excepcionalmente estable de GFP conocida como "superfolder GFP" (139). En este sistema, la proteína blanco es incorporada como una fusión N-terminal con un fragmento de GFP, GFP11 (residuos 215-230), mientras que el resto de GFP, GFP1-10 (residuos 1-214) es co-expresado desde otro vector. Por lo tanto si la proteína es expresada de forma soluble, el fragmento GFP11 puede interaccionar con GFP1-10, llevando al desarrollo de fluorescencia (138).

Otra aproximación, es la de expresar la proteína blanco como fusión N-terminal con un marcador de selección como por ejemplo cloranfenicol acteiltransferasa (CAT; 25 kDa). Se observó que cuando la proteína de fusión es expresada de forma soluble, la bacteria es resistente a mayores concentraciones de cloranfenicol que cuando se expresa de manera insoluble (140). Utilizando este método fue posible aislar variantes más solubles de la proteína humana citocromo P450(141).

Más recientemente, otro marcador de selección fue utilizado en forma partida, como para el caso de GFP, acoplándose la estabilidad de la proteína blanco con la resistencia a un antibiótico *in vivo*. La proteína se insertó como parte de una fusión tripartita entre los residuos 196-197 del gen de la TEM1-β-lactamasa (22). Cuando la proteína blanco se expresa de forma

soluble, los 2 dominios de la β lactamasa pueden interaccionar, confiriendo resistencia a antibióticos β lactámicos (22).

Finalmente, otro método desarrollado para la selección de mutantes solubles es el denominado "colony filtration blot" (CoFi-blot) (142, 143). Este método está basado en que las proteínas solubles pueden ser separadas de los CI mediante filtración a nivel de la colonia. Luego de transformar las bacterias con la librería de mutantes, las colonias son transferidas a una membrana filtrante, donde la expresión es inducida y las células lisadas. Las proteínas solubles pueden entonces, difundir a través del filtro y unirse a una membrana de nitrocelulosa para su detección mediante por ejemplo un anticuerpo anti-His (142, 143).

Así como en HTS usualmente muchas condiciones de expresión son evaluadas para la misma proteína, en evolución dirigida, una librería de mutantes es generada de forma aleatoria y analizada para la solubilización de la proteína blanco. Los puntos claves en esta estrategia son la diversificación de la librería y el método de selección empleado para encontrar aquellos mutantes que presentan mayor solubilidad.

1.10 Caracterización proteica

La obtención de la proteína recombinante en forma soluble, no asegura su correcto plegamiento. Por ejemplo, muchas veces si bien la proteína de interés se obtiene en forma soluble, se encuentra formando agregados de alto peso molecular, lo que es indicativo de regiones mal plegadas. Por esta razón es recomendable un último paso de purificación mediante cromatografía de exclusión molecular (SEC, del inglés size exclusión chromatography), no solo para eliminar algunas impurezas sino también para determinar el estado oligomérico de la proteína. Este paso también puede realizarse a nivel analítico y en forma HTS, utilizando cantidades en el orden de los microgramos, mientras aún se está evaluando las diferentes condiciones de expresión. Mediante el acoplamiento de placas de 96 pocillos conteniendo bolillas para la purificación de proteínas por IMAC (HisMultiTrap FF 96-well plates; GE Healthcare), con minicolumnas de SEC analíticas (SuperdexTM 5/150 GL; GE Healthcare) y el uso de un "autosampler", el proceso se puede realizar de manera totalmente automática (79, 144, 145).

Si bien se han obtenido importantes avances, en el campo de la producción de PRs, aún no se ha logrado generar un protocolo universal, por lo que frecuentemente es necesario realizar una extensa optimización de las condiciones de expresión. El continuo avance de nuevas metodologías para la expresión de PRs, es un factor clave en vías de aumentar las probabilidades de éxito en la obtención de un producto soluble y homogéneo.

2. Generación de proteínas de unión

2.1 Generalidades

Numerosas aplicaciones en el área de las PRs necesitan de la generación de la proteína blanco para su uso como proteínas de unión o "binders". Las proteínas de unión a su vez, pueden ser utilizadas para la purificación de otras moléculas, generación de inhibidores intracelulares, cristalización de proteínas blanco, reactivos para entender procesos biológicos complejos y/o aplicaciones clínicas diversas.

Actualmente, los AcMo son las moléculas de unión más utilizadas en el área clínica, existiendo más de 25 AcMo aprobados y comercializados para el tratamiento de diversas enfermedades en Europa y USA. Sus aplicaciones terapéuticas incluyen el tratamiento exitoso de diferentes tipos de tumores, enfermedades cardiovasculares, autoinmunes e infecciosas, así como también el control de rechazo de trasplantes de órganos entre otras (9). Estas proteínas pueden presentar una alta afinidad y especificidad contra su blanco además de poder ejercer funciones efectoras como ser citotoxicidad dependiente de anticuerpos (ADCC) y la citotoxicidad dependiente de complemento (CDC). Sin embargo, debido a que son moléculas grandes (>150 kDa), tienen poca penetración en tumores sólidos por lo que sólo un pequeño porcentaje de la dosis administrada localiza en el tumor en estos casos, disminuyendo así su citotoxicidad (146). Si bien se han desarrollado AcMo conjugados a toxinas para aumentar su efectividad terapéutica, estos también poseen complejos patrones de glicosilación así como varios puentes disulfuro, debiendo ser producidos en células de mamífero y aumentando los costos de producción (147). Además, en ciertas aplicaciones, como imagenología *in vivo*, las funciones efectoras de los AcMo pueden llevar a efectos no deseados, y su larga vida media en suero afecta negativamente el contraste ruido/señal (148).

Sin embargo, no todos los sistemas inmunes poseen moléculas con plegado tipo Ig como base para el reconocimiento de moléculas o antígenos. Por ejemplo, la Lamprea marina (*Petromyzon marinus*), evolucionó un sistema inmune adaptativo que está basado en proteínas con motivos ricos en leucina (149). Además, muchas proteínas que no necesariamente están involucradas en la inmunidad adaptativa, pueden mediar interacciones de alta especificidad y afinidad. Por lo tanto, la exploración de nuevos plegamientos proteicos o "scaffolds", para su modificación y posterior uso como nuevas proteínas de unión es una aproximación válida en este contexto (150, 151).

En este sentido, en la última década, se han generado nuevas proteínas de unión mediante el uso de la ingeniería genética y evolución dirigida, para ser aplicadas en el área básica y también áreas aplicadas de la biotecnología y la medicina (6, 152). Estas moléculas no necesariamente están basadas en el plegamiento tipo inmunoglobulina (Ig) y poseen mutaciones en su estructura que le permiten generar una superficie de unión capaz de reconocer, al igual que en el caso de los AcMo, un blanco específico. Estas nuevas proteínas de unión presentan ventajas como ser: pequeño tamaño (6-20 kDa), ausencia de modificaciones post-traduccionales, estabilidad térmica y producción en altos rendimientos en *E. coli*, disminuyendo sustancialmente los costos de producción (6, 153).

Para la generación de una proteína de unión, es necesario generar bibliotecas combinatoriales, donde se mutan sitios específicos de la superficie de la proteína de forma aleatoria. Esto se logra a nivel del ADN, al mutar aleatoriamente posiciones seleccionadas usando ya sea codones degenerados o trinucleótidos (154). De esta manera, se genera una librería conteniendo billones de moléculas diferentes con regiones constantes y regiones variables bien definidas, generándose así proteínas con superficies de unión contra prácticamente cualquier molécula (155). Posteriormente, se debe aplicar un método de selección que permita aislar los escasos clones positivos que reconocen la proteína blanco entre los billones de clones fútiles manteniendo a la vez unido el fenotipo (proteína), con el genotipo (ADN correspondiente). El proceso de selección, puede ser dividido en 3 pasos: **diversificación, selección y amplificación.** Esto incluye la expresión de una librería proteica o peptídica; enfrentamiento de las moléculas de unión con la molécula blanco y sucesivos lavados; amplificación de las moléculas seleccionadas e identificación de los clones positivos.

2.2 Diferentes "scaffolds" para la generación de proteínas de unión y sus aplicaciones.

Varios plegamientos proteicos han sido seleccionados para ser utilizados como binders para diversas proteínas blanco. Las estructuras de estos scaffolds, han sido optimizadas para que sean más robustos, tengan mayor estabilidad termodinámica y mayor rendimiento en su expresión. La optimización del scaffold es llevada a cabo normalmente por la introducción de mutaciones puntuales, donde se mutan determinados aminoácidos mediante un "diseño consenso"(156). El mismo, se basa en que los residuos más conservados, corresponden a los residuos funcionalmente importantes como resultado de una diversificación y selección durante el proceso evolutivo de la proteína. Estos residuos conservados son necesarios para el mantenimiento y proceso de plegamiento de la proteína y evitan la agregación de la misma.

Por lo tanto el reemplazo de un residuo por su correspondiente residuo consenso, mejoraría la estabilidad y/o la eficiencia de plegamiento del scaffold, haciéndolo más resistente a la introducción de mutaciones aleatorias en otros sitios (157).

Más de 50 scaffolds diferentes han sido propuestos en los últimos 20 años (152) entre los que destacamos a los Affibodies (158), DARPins (159), Anticalinas (160), Dominios Kunitz (161), Nanobodies (162) y Afitinas (11) (Figura 1).

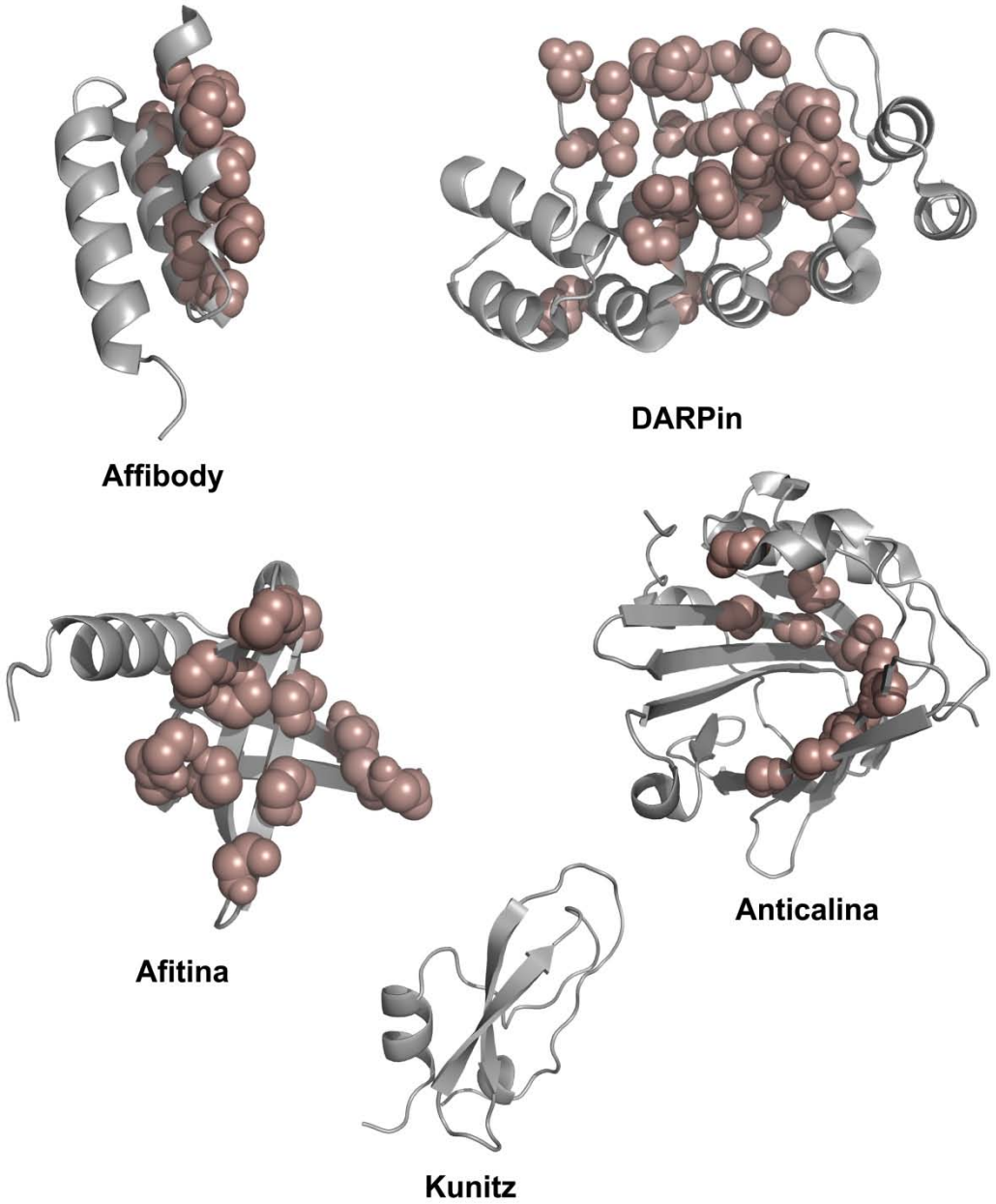


Figura 1. Estructura de los scaffolds mencionados. Se muestran las estructuras cristalográficas para Affibody (PDB: 3MZW), DARPin (PDB: 1SVX), Anticalina (PDB: 3DSZ), Afitina (PDB: 1AZP) y dominio Kunitz (PDB: 2KNT). Las esferas color salmón representan los sitios mutados aleatoreamente para formar las superficies de unión.

2.2.1 Affibodies

Los affibodies (del inglés Affibody) son proteínas derivadas del dominio B de la región de unión a inmunoglobulinas de la proteína A de *Staphylococcus aureus*. El dominio B fue modificado mediante mutaciones puntuales dando el dominio Z que presenta mayor estabilidad química, siendo ésta la base para la generación de los affibodies. Los affibodies están formados por un único dominio de 58 residuos (6.5 kDa) con tres α-hélices, no contienen cisteínas y pueden ser producidos en *E. coli* con altos rendimientos (158). Las superficies de unión son generadas al mutar de forma aleatoria 13 posiciones ubicadas en las α-hélices 1 y 2, que correspondían a la superficie de unión para el Fc de anticuerpos (163) obteniéndose afinidades en el orden picomolar (164).

Una aplicación de las moléculas de unión es la de imagen *in vivo* o radioinmunodiagnóstico (RID) y radioinmunoterapia (RIT). En este sentido, el uso de anticuerpos acoplados a radionucleidos presenta algunas desventajas como se mencionó con anterioridad. En estudios en ratones, los affibodies han mostrado tener una alta penetrancia en tumores sólidos y una rápida biodistribución y eliminación (165). Además, debido a su pequeño tamaño, los affibodies han sido producidos mediante síntesis química de péptidos lo que permite la incorporación sitio específica de varios grupos químicos como sondas fluorescentes o grupos quelantes para la unión de átomos metálicos radioactivos en un único proceso (166). En este sentido, un affibody dirigido contra el marcador tumoral, el receptor tirosina-quinasa del factor epidérmico de crecimiento 2 (HER2), fue sintetizado químicamente donde se le acopló el agente quelante DOTA al amino terminal (dando el afibody ABY-002). Este fue eficazmente marcado con ^{111}I, e inyectado en ratones, mostrando un eficiente y específico pegado en los tumores de ovario SCOV-3. De esta manera fue posible visualizar los tumores que sobreexpresaban HER2, 1 hora luego de la inyección (167). Estudios clínicos fueron realizados demostrándose que ABY-002 también reconoce HER2 en humanos por lo que podría ser utilizado en RID en el futuro (168).

Además, los affibodies han sido utilizados en otras áreas biotecnológicas como la cromatografía de afinidad con propósitos de purificación de proteínas, al unir específicamente varios blancos (169-171). Asimismo, moléculas de affibodies han sido fusionadas a diferentes enzimas como la β-galactosidasa, o la peroxidasa de rábano, permitiendo la detección colorimétrica de la proteína blanco a la que unen (172, 173). Finalmente un affibody que bloquea la DNA polimerasa, inhibiéndola a temperatura ambiente es utilizado y comercializado como “Hot start PCR reagent” (Phusion Hot Start II High Fidelity DNA Polymerase, Thermo Scientific) (Figura 1).

2.2.2 DARPins

Las DARPins (del inglés designed ankyrin repeat) son proteínas termoestables (T̊m>85 C), basadas en un módulo de motivos repetidos de anquirina (AR) de 33 residuos con 7 posiciones aleatorias. Poseen una estructura con una vuelta β seguida por dos α-hélices antiparalelas y continuadas por un loop que conecta con la vuelta β del siguiente dominio. Usualmente las DARPins contienen 3 módulos variables introducidos entre dos módulos conservados que funcionan como tapas N y C-terminales sellando el núcleo hidrofóbico (166 residuos totales, 21 posiciones aleatorias) (174). Mediante el uso de este scaffold se han podido seleccionar proteínas de unión contra una diversa gama de blancos alcanzándose afinidades del orden picomolar (175, 176). Las DARPins han sido evolucionadas para unir proteasas, quinasas y proteínas de membrana. También han sido marcadas radioactivamente con éxito, conjugadas con toxinas o fusionadas con citoquinas o proteínas citotóxicas (153). El programa más avanzado de estas moléculas en el área clínica corresponde a un inhibidor potente del factor de crecimiento vascular endotelial A (VEGF-A) (IC_{50}<10pM), actualmente en fase clínica I/II (MP0112). Estudios con pacientes que sufrían de edema diabético macular, mostraron que esta molécula es segura y bien tolerada como inyección intravitreal (177).

Recientemente, se logró también redirigir partículas de adenovirus del serotipo 5 a células que expresan diferentes marcadores tumorales utilizando DARPins bi-especificas (fusión de 2 DARPins con especificidad a 2 blancos diferentes o sitios diferentes) (178). En este diseño, una de las DARPins (1D3) reconoce la fibra “knob” del adenovirus mientras que la otra DARPin puede reconocer alguno de los diferentes marcadores tumorales pudiendo dirigir dichas partículas hacia diferentes tipos celulares de manera específica (178).

Por otro lado, una nueva generación de DARPins fue diseñada recientemente conocida como loopDARPins. Estas, contienen un loop extendido de 19 residuos (con 10 posiciones aleatorias) en el giro β que protruye en uno de los AR (176). Lo sorprendente de este nuevo diseño, es que se logró obtener mediante una única ronda de ribosome display, proteínas de unión con afinidades de hasta 30 pM contra diferentes miembros de la familia de reguladores antiapoptóticos BCL-2 (BCL-2, BCL-XL, BCL-W y MCL-1). Además en algunos casos las proteínas de unión aisladas no presentaron reconocimiento cruzado entre los diferentes miembros de la familia a pesar de que presentan una alta homología estructural (176). Por último, este loop podría estar dirigiendo en parte, la unión de las loopDARPins hacia estructuras tipo bolsillo o profundas como es el de muchos sitios activos, ya que todos los loopDARPins seleccionados en este trabajo unen en o cerca del bolsillo de interacción de los distintos miembros de la familia BCL-2 (176).

2.2.3 Anticalinas

Las anticalinas (160-180 residuos), son derivadas de las lipocalinas y constituyen una familia de proteínas presentes en humanos insectos y otros organismos. Están formadas por un barril β con 8 hebras antiparalelas unidas a través de 4 loops estructuralmente variables que unen naturalmente ligandos pequeños (179). Utilizando este "scaffold" se han podido obtener proteínas de unión con alta especificidad y afinidad tanto para compuestos pequeños como haptenos así como también proteínas al mutar de forma aleatoria 16-24 aminoácidos en los loops de interacción natural (148, 160, 179). Estas proteínas son producidas eficientemente en *E. coli* y levaduras, y varias anticalinas fueron seleccionadas contra blancos terapéuticos. Un ejemplo es la anticalina PRS-050 (Pieris) que reconoce con afinidades en el orden nanomolar al VEGF-A. Esta anticalina está actualmente en fase I de estudios clínicos con pacientes con tumores sólidos avanzados, mostrando ser bien tolerada (180). En estos estudios, la anticalina fue fusionada a una molécula de 40 kDa de polietinolglicol (PEG) para aumentar su vida media plasmática en hasta 6 días (180). En otro trabajo, se generó una anticalina capaz de unir radionucleidos en complejo con el agente quelante, ácido dietilenetriamina pentaacético (DTPA). Esta anticalina une al DTPA acomplejado con isotopos de uso común en RIT y RID como ser iones lantánidos Y^{3+} (aplicado en radioterapia o tomografía de emisión de positrones) Tb^{3+} (luminiscencia de lantánidos), Gd^{3+} (útil en imagen por resonancia magnética) y Lu^{3+} (aplicaciones tanto en radioterapia como imagenología) (181). Por lo tanto, la misma puede ser una herramienta muy útil al fusionarla con otra molécula que dirija el complejo anticalina-Me-DTPA a por ejemplo células tumorales (148).

2.2.4 Dominios kunitz

Los dominios Kunitz, son inhibidores reversibles de serina proteasas típicamente de origen humano, contienen 58 residuos, 3 puentes disulfuro y 3 loops que pueden ser mutados sin desestabilizar su estructura. Al randomizar 9 posiciones en sus loops se lograron seleccionar mediante "phage display" inhibidores potentes de la calicreína plasmática humana alcanzando una constante de inhibición de 15 pM (161). La calicreína es una serina proteasa plasmática, que juega roles importantes en la inflamación y coagulación. También está involucrada en enfermedades como el angioedema hereditario, debido a una deficiencia genética en el inhibidor C1-esterasa, que inhibe a la calicreína y el factor de coagulación XIIa (182). La droga Ecallantide (DX-88), es un derivado de dominios Kunitz y surge como un inhibidor de alta afinidad para la calicreína *in vivo* (182). Este es el primer scaffold en alcanzar el mercado al ser aprobado para el tratamiento del angioedema hereditario en diciembre de 2009, y es comercializado actualmente por Dyax bajo el nombre de Kalibitor® (6, 153).

2.2.5 Nanobodies

Los camélidos, además de producir AcMo en un formato típico (2 cadenas pesadas + 2 cadenas livianas), producen un formato atípico que posee sólo las cadenas pesadas y además carecen del dominio CH1. El dominio variable de reconocimiento antigénico proveniente de estas cadenas pesadas es llamado VHH (162). El sitio de unión al antígeno estaría formado en este caso por sólo 3 loops hipervariables (H1, H2 y H3), donde el loop H3 es en promedio más largo que el de los anticuerpos convencionales. Esto puede llevar a una preferencia de unión hacia estructuras tipo bolsillo o superficies cóncavas en las proteínas como ser sitios activos, como fue demostrado con diferentes complejos VHH-lisozima (10). El dominio VHH puede expresarse sin las regiones constantes y se conoce como nanobody. Estos son termoestables (60-80˚C) y pueden ser producidos con altos rendimientos en el periplasma de *E. coli* (183). La estructura de los nanobodies, consiste en el típico plegamiento Ig, con una hoja β de 4 hebras y otra hoja β de 5 hebras conectadas por loops y un puente disulfuro (Cys23-Cys94) empaquetado contra un triptófano conservado (162).

Los nanobodies son generalmente obtenidos luego de inmunizar un camélido, seguido del clonado del repertorio de genes V de los linfocitos B de sangre periférica y posterior selección de nanobodies con alta afinidad mediante phage display (Kd en el orden de nano y picomolar) (184). La fusión génica de nanobodies con una proteína fluorescente, conocidos como cromobodies o fluorobodies, han permitido rastrear diferentes antígenos en varios compartimentos celulares *in vivo* (185, 186). En otro trabajo, un nanobody fue conjugado con nanopartículas de oro, lo que puede utilizarse para dirigir terapias fototérmicas tras iluminación con luz láser (187). También se lograron desarrollar nanobodies biespecíficos contra toxinas de escorpión (toxinas Aahl´y Aahll) encontrándose un efecto protector en ratones donde se inyectaron dosis letales del veneno (188). Finalmente varios nanobodies se encuentran en fases clínicas I y II contra diversas patologías como ser la Artritis reumatoidea y el virus respiratorio sincicial entre otros, mostrando un futuro prometedor en el área médica (189).

2.2.6 Afitinas

Las afitinas (del inglés “affitins”), son derivadas de la proteína de unión al ADN, Sac7d de la arquea *Sulfolobus acidocaldarius*, perteneciente al grupo de proteínas acidofílicas e hipertermostables. Sac7d es una proteína pequeña de 66 residuos (7 kDa), y no presenta puentes disulfuro. Posee un plegamiento de unión a nucleótidos/oligosacáridos y un barril β incompleto de cinco hebras tapado por una α-hélice C-terminal. Sac7d, es expresada con altos rendimientos en *E. coli,* es estable entre pH 0-12, temperaturas de hasta 90̊C, y presentó

estabilidad química como ser 8M urea y detergentes (190). Librerías conteniendo 14 sitios mutados de forma aleatoria, fueron diseñadas basándose en las estructuras cristalográficas disponibles. De esta forma se logró obtener proteínas de unión con afinidades en el orden picomolar para la proteína bacteriana PulD, que al ser exportadas al periplasma podían inhibir su oligomerización *in vivo*, bloqueando la vía de secreción tipo II (11). Así se cambió una proteína de unión al ADN a una de unión a proteínas. Además, la fusión entre la afitina de unión a PulD con la proteína roja fluorescente (mCherry), mostró ser estable y pudo ser utilizada como agente de inmunolocalización para PulD (191). También se obtuvieron afitinas contra IgG_1 humana (utilizando librerías de 10 y 14 sitios randomizados), con afinidades en el rango nanomolar. A pesar de las mutaciones presentes, estas afitinas tenían estabilidad térmica (hasta 80ºC) y química (pH 0 a 12) (192), haciéndolo un candidato interesante para explorar en nuevas aplicaciones.

Luego de todos los ejemplos citados, queda establecida la importancia de estas moléculas optimizadas para su función de "binders" en las distintas aplicaciones dentro de las áreas de la biotecnología y biomedicina. Muchas de las aplicaciones para éstas moléculas se resumen en la Figura 2.

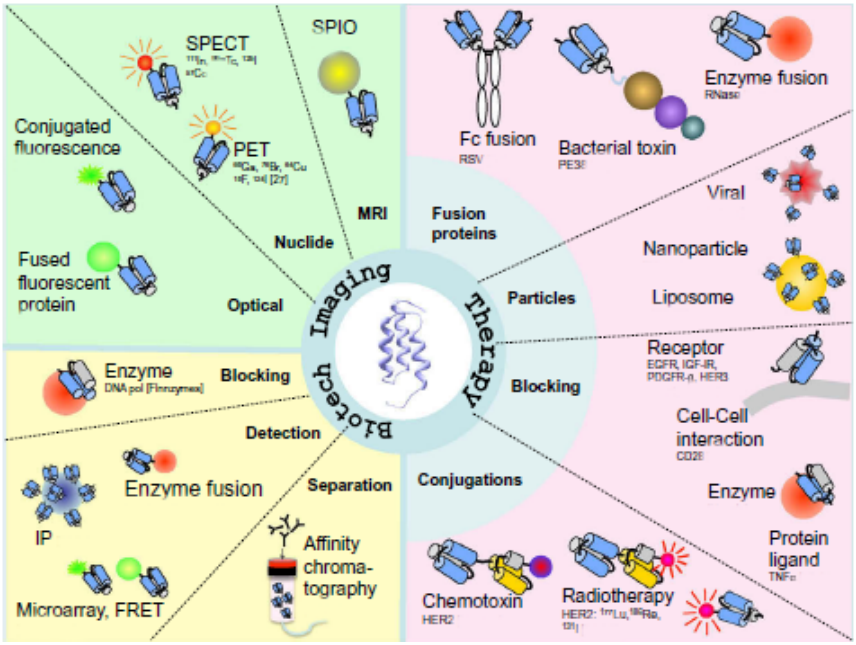


Figura 2. Esquema de las aplicaciones en las que han estado involucrados los affibodies. Adaptable también a los diferentes scaffolds mencionados. Adaptado de (158).

2.3 Diferentes métodos de selección para la obtención de "binders"

La era de las librerías combinatoriales, se inició a mediados de los 80 con la invención de la técnica de selección conocida como "phage display" por G.P. Smith (193). En años posteriores, han sido desarrollados muchas variaciones de métodos de muestreo o "display" de proteínas y péptidos basados en librerías combinatoriales aleatorias. Las diferentes metodologías de display pueden ser divididas en dos principales categorías: métodos dependientes de células y métodos independientes de células.

En los métodos dependientes de células, las proteínas de la librería son presentadas en la superficie de fagos, de células o expresadas en compartimentos celulares. Esto tiene como desventaja que la diversidad posible de la librería está limitada por la eficiencia de transformación de las células. Dentro de este tipo de métodos encontramos: "Phage display" (194), "Yeast display" (195) y "*E. coli* surface display" (196), siendo phage display el método más utilizado.

En los métodos independientes de células, la transcripción y traducción se realizan *in vitro*, por lo que al no estar limitado por la eficiencia de transformación, la diversidad de las librerías evaluadas es mayor (pudiendo alcanzar 10^{13} moléculas diferentes). Además, al realizarse totalmente *in vitro*, se pueden introducir mutaciones en las etapas de amplificación (mediante PCR proclive al error), dando lugar a un proceso de verdadera evolución Darwiniana, en contraposición con la selección de una librería constante ya existente (197). Este tipo de metodologías además permite realizar las selecciones en diferentes condiciones de pH y temperatura siempre y cuando el vínculo genotipo-fenotipo sea lo suficientemente robusta (198). Dentro de esta categoría se destacan Ribosome display (199) y mRNA/cDNA display (200, 201).

2.3.1 Phage display

Phage display es el método más utilizado para la selección de proteínas de unión. En este método, los péptidos o proteínas son presentados en la superficie de las partículas de bacteriófagos filamentosos (más comúnmente el fago M13) que contienen el ADN codificante en su interior, manteniendo así el vínculo físico entre fenotipo y genotipo. Para la presentación de las moléculas de unión en la superficie del fago, la librería de mutantes es fusionada a alguno de los genes de la cubierta del virus como ser pVIII (producido en 2700 copias, usualmente para la presentación de péptidos) o pIII (producido en 5 copias, usualmente para la presentación de proteínas) (194, 202). Ambas proteínas poseen péptidos señal que dirigen la fusión al periplasma de la bacteria. Una vez en el periplasma, la peptidasa líder Lep de *E. coli*, remueve el péptido señal creando la forma madura de la proteína de la capside.

Recientemente se han sugerido las proteínas pVII y pIX como alternativas a pIII para la presentación de proteínas de unión (203). Las células de *E. coli* conteniendo los plásmidos para las fusiones mencionadas, son infectadas con un fago auxiliar que aporta todas las proteínas virales necesarias para el ensamblado de las partículas del fago que presentarán las diferentes moléculas de unión en su superficie. Una vez producidos los fagos, estos son incubados por ejemplo en una placa conteniendo el blanco inmovilizado para realizar la selección. Los fagos que no se unen al blanco son lavados, mientras que los que permanecen unidos son luego eluídos mediante una corta incubación a pH bajos, o corte proteolítico. Los fagos eluídos son utilizados para infectar nuevas células de *E. coli* para amplificar los clones seleccionados y las partículas de fago son rescatadas por superinfección con el fago auxiliar, creando una nueva librería de fagos que puede ser utilizada en una nueva ronda de selección (usualmente 3-5 rondas son necesarias para obtener variantes de alta afinidad) (155) (Figura 3).

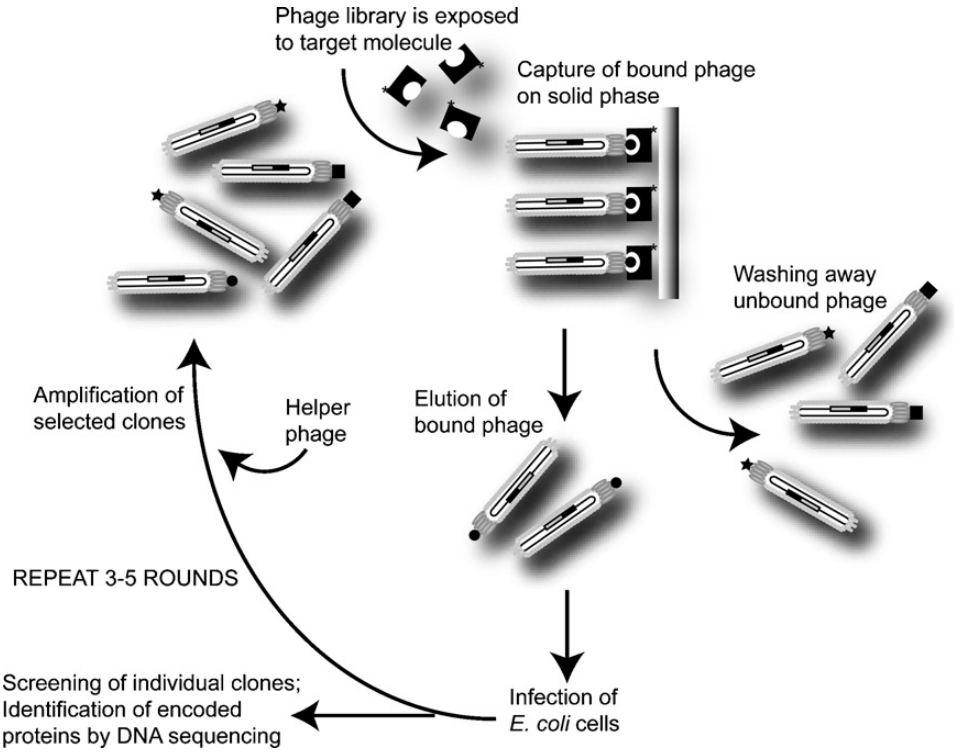


Figura 3. Representación esquemática de la selección de moléculas de unión mediante phage display. Adaptado de (155).

2.3.2 Ribosome display

Esta metodología es la más exitosa dentro de los métodos independientes de células, descripta por primera vez para péptidos de unión en 1994 por Mattheakis et al. (204) y modificada en 1997 para proteínas de unión por Hanes et al. (199). La misma consiste en el acoplamiento no covalente de un complejo ternario formado por el ARNm codificante, el polipéptido saliente y

el complejo ribosomal. El ADN correspondiente a la librería de mutantes es transcripto (generalmente con la polimerasa T7), y las moléculas de ARNm obtenidas son traducidas *in vitro* con una cantidad estequiométrica de ribosomas, provenientes por ejemplo del extracto S30 de *E. coli*. La formación estable del complejo ternario se da gracias a la ausencia de codón stop en el ARNm. Esto provoca que el último aminoácido del polipéptido saliente quede conectado al peptidil-tRNA ya que no puede liberarse del ribosoma debido a que los factores de liberación interaccionan con los codones stop. Además, la liberación del ARNm ocurre solo después de que la proteína sintetizada y el ARNt son liberados, reacción catalizada por el factor de reciclaje ribosomal (205). Por lo tanto, los ribosomas que traducen ARNm sin codones stop, quedan atrapados en una forma en que la proteína emerge del ribosoma y el ARNm está todavía unido al mismo, conectando así el fenotipo con el genotipo.

Para que la proteína quede plegada en su forma correcta mientras está unida al ribosoma, es que se introduce una secuencia espaciadora en la región carboxilo terminal que se encuentra codificada como una fusión 3' en la librería de ADN. Esta secuencia extra, ocupa el túnel ribosomal, y provee de la flexibilidad necesaria para el plegamiento de la proteína como una unidad independiente y para la unión con la proteína blanco (205).

Finalmente, una vez formados lo complejos ternarios, estos son sometidos a la interacción con la proteína blanco inmovilizada a una superficie (Figura 2). La selección se realiza a bajas temperaturas y altas concentraciones de Mg^{2+} con el objetivo de estabilizar los complejos ternarios. Las proteínas de unión con poca o ninguna afinidad son lavadas, mientras que las que presentan unión contra la proteína blanco, pueden ser recuperadas. Esto último se logra con la adición de EDTA, que produce la disociación de las subunidades chica y grande del ribosoma, con la consecuente liberación del ARNm. De esta manera la información genética de las proteínas de unión puede ser amplificada y analizada. Repitiendo este proceso durante 3-6 rondas es posible obtener proteínas de unión de alta afinidad (en el orden de nM-pM) (205-207) (Figura 4). Una ventaja de RD, es que no es necesario romper la interacción entre la proteína blanco y el clon de la librería que se encuentra unido a ella, para la obtención de la información genética. De esta manera no es más difícil aislar secuencias correspondientes a complejos de muy alta afinidad (que serían difíciles de disociar) con respecto a los de menor afinidad.

Para la obtención de proteínas de alta afinidad, que presenten un componente de disociación muy bajo, una estrategia es realizar lo que se conoce como "off rate selection". La misma consiste en exponer la librería producida, con la proteína blanco biotinilada e inmovilizada. En

un segundo paso se adiciona mas proteína blanco pero sin biotinilar, permitiéndose entonces que aquellos mutantes con una disociación rápida se unan a ellas. Finalmente se procede a un lavado exhaustivo en donde las variantes con bajas constantes de disociación son recuperadas tras la elución (208).

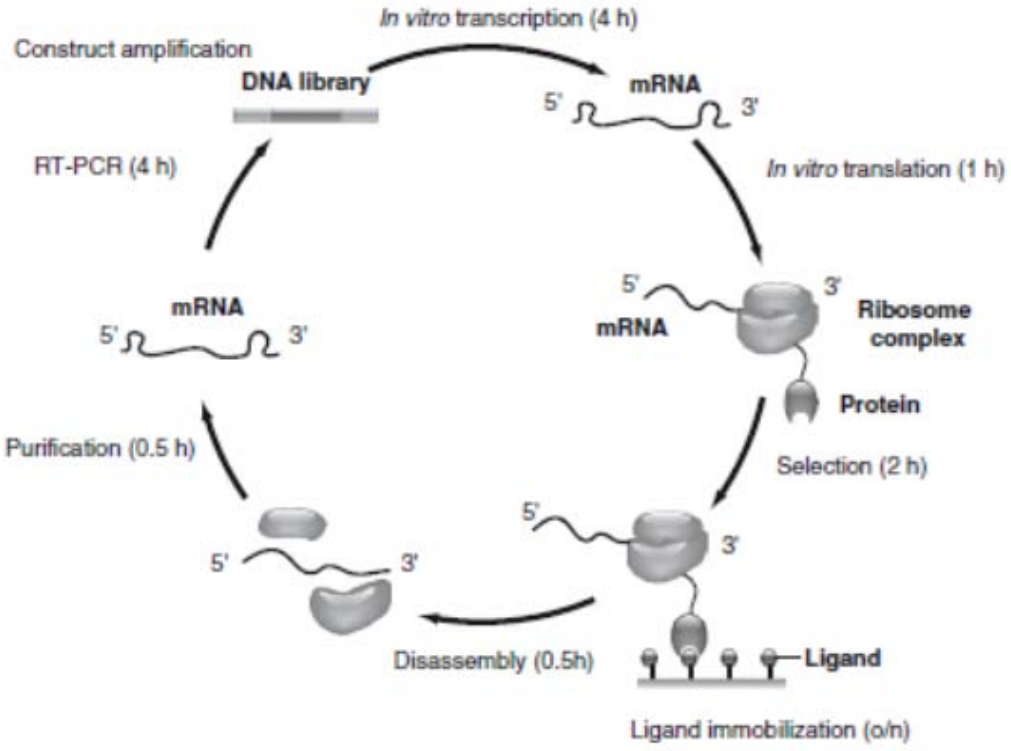


Figura 4. Representación esquemática del proceso de selección por RD. Adaptado de (209).

Más recientemente, la metodología de RD fue optimizada al utilizar componentes para la traducción *in vitro* purificados (PURE system), en lugar del extracto S30 de *E. coli*. Este sistema está compuesto por los factores y enzimas responsables de la expresión génica en *E. coli.* Debido a que los mismos son purificados por IMAC, no contiene nucleasas y proteasas en comparación con los extractos S30 de *E. coli* (209, 210). Utilizando este sistema en RD, se logró realizar la selección incluso en presencia de codón stop, ya que los factores de liberación no se encuentran presentes, lo que ayuda a detener el ribosoma y estabiliza el complejo ternario (210-212). Incluso se observó que luego de incubar los complejos ternarios por 1 hora a 50°C, la eficiencia de selección se vio reducida por un factor menor a 10 (212). Además, otro elemento que se adicionó a las librerías cuando se utilizó este sistema es la secuencia secM de *E. coli* hacia el extremo 3'. Esta secuencia (FSTPVWISQAQGIRAGP) es parte del mecanismo de detención de la traducción, y es capaz de interaccionar con residuos presentes en la salida del túnel ribosomal, estabilizando la interacción proteína-ribosoma (213).

Finalmente, se diseño otra estrategia para generar un protocolo de RD altamente estabilizado, muy útil para la selección principalmente de péptidos de unión. En este sistema, el ARNm codifica hacia el 5'-UTR, un motivo de RNA llamado cv seguido de la librería de péptidos en fusión con la proteína de asociación al RNA Cv (Cvap). De esta manera, la proteína Cvap una vez traducida, interacciona fuertemente con el motivo de ARN presente hacia el 5'-UTR de la misma molécula de ARN que la codifica, conectando de esta manera el fenotipo con el genotipo (214).

2.3.3 mRNA/cDNA display

Al igual que ribosome display, mRNA display utiliza el complejo formado entre el ARNm y el polipéptido codificado por ese ARNm como estrategia de selección. La diferencia con el ribosome display radica en la formación de un enlace covalente entre el ARNm y el polipéptido a través del antibiótico puromicina, generándose así una mayor estabilidad del complejo (205).

La metodología de cDNA display surge como una optimización del método de mRNA display, en donde el complejo ARNm-proteína es rápidamente convertido a cDNA-proteína, siendo el cDNA el que se une covalentemente a la proteína mediante un espaciador o "linker" con puromicina (201, 215). Además en este método la unión del linker de ADN conteniendo a la puromicina con el ARN se realiza utilizando la RNAsa T4, siendo este proceso más eficiente. Tras la traducción *in vitro*, en fase sólida, la proteína sintetizada, queda fusionada a la puromicina y el ARNm es retrotranscripto usando la región cebadora en el linker, por lo que el cDNA generado forma parte covalente del linker Puromicina-ADN (Figura 5). La molécula formada por ARN-cDNA-Proteína es liberada de la superficie solida tras el clivaje con una enzima de restricción específica para la región de ADN doble hebra presente en el linker y puede ser utilizada para las rondas de selección de proteínas de unión (Figura 5) (201).

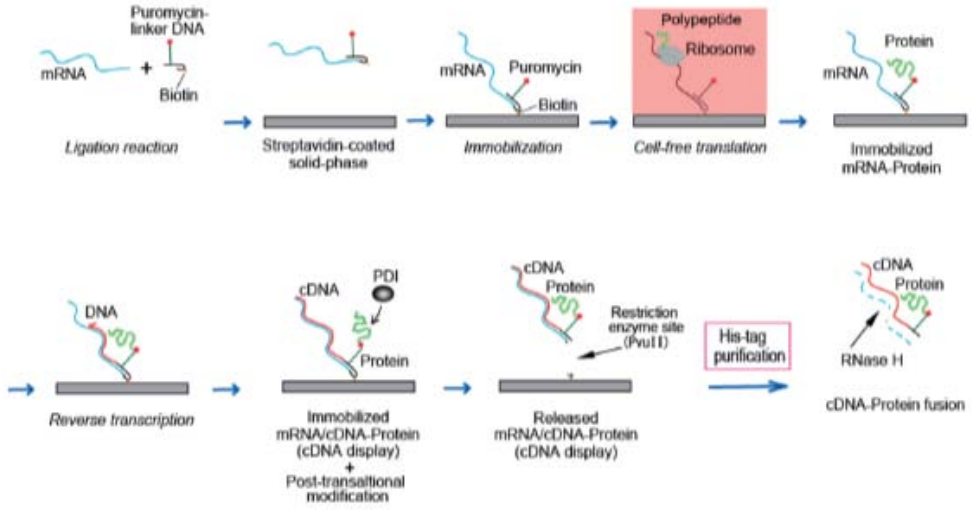


Figura 5. Esquema de las etapas involucradas en el desarrollo de cDNA display. Adaptado de (201).

3. Cristalogénesis: nuevas estrategias para un viejo problema

3.1 Generalidades

La obtención de las estructuras tridimensionlaes de las proteínas de unión así como también de los complejos con sus blancos ha sido fundamental para poder mejorar las funciones y aplicaciones de los diferentes "scaffolds". La información estructural, puede utilizarse para determinar cómo los diseños de las librerías pueden ser mejorados y qué factores gobiernan el reconocimiento molecular en estos sistemas (216). Por otro lado, las estructuras cristalográficas han además permitido entender diferentes procesos biológicos así como también diseñar pequeñas moléculas para la inhibición de la actividad de blancos terapéuticos que se utilizan en tratamientos clínicos (7) entre otras aplicaciones.

Actualmente, la cristalografía de rayos X, es la metodología que contribuye con mayor número de estructuras de alta resolución de macromoléculas biológicas en la base de datos del PDB (Protein Data Bank).

A pesar de la importancia de las estructuras cristalográficas y el impresionante avance en estos 100 años en la difracción de rayos X, al día de hoy la obtención de muchas de las estructuras de macromoléculas se encuentra con 2 grandes obstáculos. Estos son: **i) la obtención de las proteínas recombinantes en forma soluble, pura y en cantidad suficiente y ii) la generación de cristales de buena calidad y con alto poder de difracción.** Esto último, se debe principalmente a que el proceso de cristalogénesis de macromoléculas biológicas es un proceso cinéticamente desfavorable. En el proceso de cristalogénesis, las moléculas disueltas

que pueden moverse libremente en solución acuosa reducen su entropía, al ordenarse de forma periódica en una red cristalina tridimensional en parte estática, formada por contactos moleculares débiles. Esta pérdida de libertad traslacional y rotacional de las proteínas representa una barrera energética en términos de energía libre, para la formación del cristal (217, 218). Sin embargo, la liberación de moléculas ordenadas de solvente de las superficies involucradas en los contactos cristalinos, compensaría esta pérdida entrópica proveyendo en última instancia la fuerza conductora para el crecimiento del cristal (219).

En vías de favorecer la formación de contactos cristalinos, se recurre a la reducción de la solubilidad de la macromolécula, alcanzando un estado de supersaturación. Esto se logra, por ejemplo al incrementar la concentración de sales y/o polímeros, haciendo que la macromolécula tienda a agregar y precipitar o en el mejor de los casos, formar puntos de nucleación que pueden derivar en la formación de cristales tridimensionales (220). Por lo tanto, para encontrar la condición de cristalización óptima para una proteína, deben evaluarse cientos o incluso miles de condiciones de cristalización diferentes hasta encontrar la composición de agentes precipitantes, tampones y aditivos en la cual la proteína cristaliza.

Al igual que para la expresión de proteínas recombinantes (ver sección 1 de la introducción), importantes avances en el campo de la automatización se han dirigido a la evaluación de condiciones de cristalización. Al día de hoy, existen tecnologías que permiten la evaluación de cientos de condiciones de cristalización de forma rápida y automática, que utilizan además cantidades mínimas de proteína, al dispensar volúmenes en el orden de los nanolitros. Adicionalmente se ha automatizado la inspección de estas condiciones cristalográficas, permitiendo su reproducción exacta así como la visualización y clasificación de los posibles "hits" en forma periódica y automática (CrystalScoreTM) (14). Si bien estos avances han permitido la cristalización de numerosas proteínas blanco, hay muchos casos en donde aún no se ha tenido éxito (incluso luego de evaluar miles condiciones de cristalización para una única proteína), por lo que nuevas estrategias son necesarias. Cuando una proteína blanco falla en cristalizar, puede ser más fructífero expandir los ensayos de cristalogénesis incluyendo variaciones de la proteína blanco, en lugar de continuar evaluando diferentes condiciones de cristalización de forma exhaustiva.

Mediante el uso de la ingeniería de proteínas (fusión con otras proteínas, mutaciones puntuales, versiones truncadas, proteínas de unión para inmovilizar regiones flexibles, remoción de sitios glicosilados, entre otras), se han podido obtener cristales y la posterior

determinación de las estructuras correspondientes que de otra forma no era posible (220-222).

3.2 Determinación de dominios o fragmentos proteicos para su cristalización

Los extremos amino y carboxilo terminales de las proteínas, son frecuentemente flexibles y desestructurados, creando un potencial impedimento entrópico para la cristalogénesis. Una manera de superar este obstáculo es el de proceder a la proteólisis limitada para "recortar" dichos extremos, generando un núcleo rígido de la proteína blanco (223). Sin embargo esto puede llevar a una heterogeneidad en la muestra debido a proteólisis incompleta. Por otro lado, las proteínas eucariotas suelen tener múltiples dominios conectados por una secuencia desestructurada, lo que puede generar gran flexibilidad, dificultando el proceso de cristalización. Por lo tanto, una estrategia puede ser la de identificar un mínimo fragmento funcional o dominio de la proteína blanco para su cristalización. Diferentes dominios pueden ser identificados mediante un análisis de secuencia y comparación con proteínas homólogas utilizando programas especializados en el diseño de construcciones truncadas. Sin embargo, muchas veces es difícil determinar los extremos de los dominios debido a que son regiones menos conservadas (224, 225). La situación se vuelve aún más compleja para los blancos en los que no hay secuencias y/o estructuras homólogas. Es así que se desarrollaron métodos experimentales para la selección de variantes "recortadas" de la proteína blanco. En este sentido, se generan librerías con deleciones incrementales en forma aleatoria para ambos extremos del gen, acoplado a un método de selección HTS de los clones que expresan las construcciones solubles (Figura 5)(226, 227). Una vez seleccionados los clones solubles, estas nuevas construcciones son utilizadas en un cribado de cristalogénesis, donde debido a ser más rígidas que la proteína completa tienen más probabilidades de cristalizar (226, 227). Las deleciones aleatorias en ambos extremos se realizan utilizando la exonucleasa III y la nucleasa de judía mung aplicando distintos tiempos de incubación. Los genes son expresados en marco de lectura con un HisTag aminoterminal y un péptido aceptor de biotina (avitag) carboxilo terminal (Figura 6). La expresión de las diferentes colonias es evaluada mediante "colony blot" con estreptavidina fluorescente para detectar el estado de biotinilación C-terminal como indicador de solubilidad y un anticuerpo anti-His para detectar el HisTag N-terminal. De esta manera se determina que ambos extremos estén presentes y por ende que no haya degradación parcial de la construcción expresada determinándose así los fragmentos mínimos a utilizar en posteriores ensayos de expresión y cristalización (226, 227).

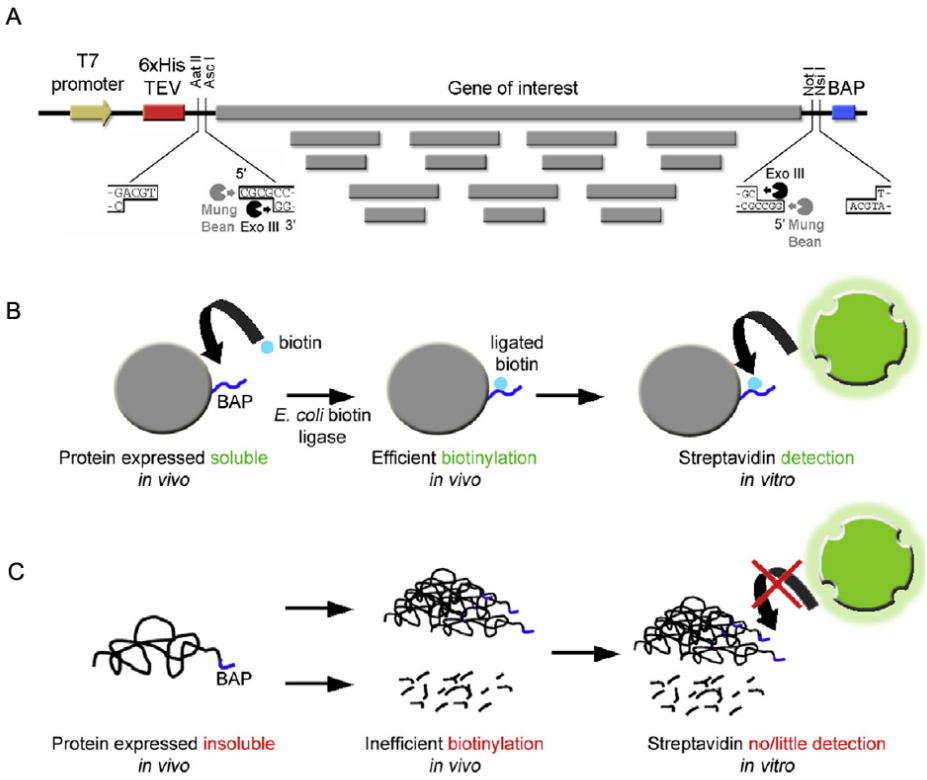


Figura 6. Esquema del método de deleción y selección. A) Deleción bi-direccional. B) y C) selección de las diferentes variantes mediante unión a la streptavidina. BAP, biotin aceptor peptide. Adaptado de (227).

3.3 Remoción de sitios glicosilados

La N- y O- glicosilación constituyen modificaciones post-traduccionales comunes, principalmente en proteínas eucariotas asociadas a la membrana, secretadas y lisosomales. Las estructuras de algunas glicoproteínas han sido resueltas donde los grupos carbohidrato se encontraron ordenados (228). Sin embargo en términos generales, la flexibilidad y heterogeneidad de los grupos carbohidrato puede significar una fracción importante del área de superficie de la proteína y por lo tanto afectar negativamente el proceso de cristalogénesis. Cuando la proteína blanco es expresada en *E. coli,* no se introducen estas modificaciones evitándose este problema. Sin embargo algunas proteínas sólo son expresadas de forma soluble en sistemas eucariotas, que son capaces de introducir diferentes motivos azúcares en la proteína. En estos casos, es posible eliminar alguno de los sitios de glicosilación en la proteína blanco al mutar la asparagina del motivo de N-glicosilación (Asn-X-Thr/Ser) a aspartato o glutamina. De manera similar con las serinas o treoninas en los sitios de O-glicosilación que pueden ser mutados a otros residuos (222). Finalmente los azúcares también pueden ser eliminados enzimáticamente con el uso de glicosidasas específicas, que cortan motivos de azúcar particulares.

3.4 Reducción de la entropía de superficie (SER)

Como ya se mencionó, la cristalización de proteínas es principalmente un proceso dirigido por la entropía, por lo que la presencia de residuos móviles en las superficies proteicas puede afectar negativamente la cristalogénesis. En este sentido, la SER (del inglés "Surface-Entropy Reduction") intenta reducir la movilidad de las cadenas laterales largas e hidrofílicas de la proteína blanco expuestas al solvente, a fin de favorecer energéticamente la formación de cristales. Los movimientos aleatorios de estas cadenas pueden incrementar la entropía del sistema al punto de inhibir la formación de la red cristalina (229). Al sustituir los parches de residuos móviles altamente entrópicos (como ser Lys, Glu y Gln) presentes en la superficie, por residuos pequeños no polares (por ejemplo alanina) se estaría disminuyendo la entropía total del sistema. Cabe resaltar, que si bien numerosas estructuras han sido obtenidas con esta aproximación (222), cambiar las propiedades electrostáticas de la proteína puede alterar su comportamiento bioquímico y su interacción con otras proteínas. Además, al cambiar residuos hidrofílicos expuestos a la superficie por residuos no polares se puede disminuir la solubilidad y estabilidad de la proteína. Finalmente si la estructura no es conocida y no hay estructura homóloga disponible, determinar los residuos apropiados a mutar sin afectar la estructura o fisiología de la proteína puede resultar difícil. En este sentido se desarrolló un programa que asiste en el diseño de las variantes a cristalizar basándose en la secuencia aminoacídica para determinar los sitios correctos a mutar (SERp Server) (230). El servidor realiza con la secuencia introducida, una predicción de los sitios que pueden contener residuos poco conservados evolutivamente, con gran entropía conformacional y que pueden estar expuestos al solvente, sugiriendo grupos de residuos a mutar.

Por otro lado, la ingeniería de superficies proteicas mediante modificación química ofrece otra aproximación a SER, presentando como ventaja que sólo los residuos expuestos en la superficie y accesibles son modificados. Sin embargo, al no ser dirigida la modificación, se modifican también residuos que pueden ser biológicamente importantes. La aproximación más común para la modificación química es la metilación reductiva de grupos amino libres (residuos de lisina y el extremo N-terminal), donde las aminas primarias son modificadas a aminas terciarias, mediante la incubación con un complejo dimetilamina-borano (231). Para el caso de residuos de lisina, esto genera lisinas dimetiladas, que son más rígidas e hidrofóbicas pudiendo además favorecer contactos proteína-proteína. Tal modificación se ha visto que disminuye la solubilidad de la proteína así como también su punto isoeléctrico. El uso de esta aproximación permitió obtener la estructura de proteínas que eran refractarias a la cristalización (231). Finalmente, esta aproximación puede complementar a SER, donde en una

primera instancia acoplando la modificación química con espectrometría de masa se podría determinar experimentalmente las lisinas que se encuentran accesibles al solvente. En un segundo paso estas se modifican a nivel genético mediante SER y se realizan los ensayos de cristalogénesis. Como ventaja de acoplar ambas estrategias está el hecho de que es posible seleccionar cuáles lisinas modificar en el segundo paso, evitándose de esta manera la modificación por ejemplo de lisinas biológicamente importantes.

3.5 Simetrización sintética de proteínas

Los contactos cristalinos consisten en interacciones intercadena o intermoleculares que se dan como resultado del proceso de cristalización de la proteína. Se ha visto que proteínas simétricas, como ser los homodímeros, son en promedio más proclives a cristalizar que proteínas monoméricas. Al formarse homooligómeros rotacionalmente simétricos, alguno de los contactos cristalinos que forman parte del cristal pueden ser los contactos ya existentes entre las diferentes moléculas, por lo que menos contactos fortuitos son necesarios (232). Bajo esta consigna surge la simetrización sintética. La misma consiste en la introducción en la superficie de la proteína de motivos que llevan a una asociación simétrica entre las proteínas, formándose un homodímero. Esto se logró por ejemplo, al generar mutantes de lisozima con una mutación a cisteína en su superficie (233). La formación de un puente disulfuro entre monómeros, permitió la obtención de 6 nuevas formas cristalinas para la lisozima que no eran posibles de obtener con el monómero. Más aún en los 6 casos los dímeros de lisozima fueron dímeros simétricos y en 2 de esos casos la simetría interna del dímero fué utilizada para construir la simetría completa del cristal. Además, al realizar un cribado de cristalización de 384 condiciones con uno de los dímeros, se encontraron cristales en 98 condiciones mientras que sólo se obtuvieron cristales en 14 condiciones con la versión nativa de la lisozima (233). Posteriormente, esta estrategia se pudo adaptar a una proteína de estructura desconocida, como fue el caso de la endogluconasa CelA de *Thermotoga marítima,* que no había sido posible cristalizar. Se seleccionaron sitios con residuos altamente polares y de baja conservación, evaluándose 8 posiciones diferentes para su mutación a cisteína y posterior formación de un S-S entre monómeros. De esta forma, uno de los mutantes permitió resolver la estructura de CelA a 2.4 Å de resolución (234).

En otra aproximación, en lugar de agregar cisteínas, se introdujeron leucinas para la formación de los homodímeros. De esta manera, se logró cristalizar y obtener la estructura de la ribonucleasa pancreática I humana (RNasaI) de la cual no había sido posible obtener cristales previamente. En este caso, se introdujeron mutaciones puntuales de 2, 3 y 4 residuos de leucina en una α-hélice de la RNasaI, para formar una interface de cremallera de leucinas que

favorezca la homodimerización (235). Los sitios para la incorporación de las leucinas se determinaron utilizando la estructura de una proteína con alta similitud como ser la RNasa A bovina. Las estructuras obtenidas de las diferentes variantes de RNasaI revelaron que las leucinas introducidas realizaban contactos hidrofóbicos con las leucinas de una segunda molécula, necesarios para el empaquetamiento cristalino. Por lo tanto la incorporación de leucinas puede promover la cristalización de una proteína si una hélice se encuentra disponible para su modificación (235).

Finalmente, en otra aproximación la simetrización sintética se logró mediante la coordinación de metales. Residuos de histidina o cisteína fueron introducidos en la superficie de MBP o lisozima formando contactos cristalinos o ensamblados oligoméricos tras la coordinación de diferentes metales (Cu^{2+}, Ni^{2+} o Zn^{2+}). Mediante esta estrategia se obtuvieron 8 estructuras nuevas de lisozima y 8 estructuras nuevas para MBP (236). Una diferencia de esta aproximación con la formación de S-S o cremallera de leucina es que pueden formarse oligómeros distintos de dímeros. Además los metales presentes en el cristal pueden proveer de las fases experimentales necesarias para la determinación de la estructura. Para el caso de la lisozima, tras agregar 2 o 4 histidinas se pudieron obtener estructuras formando dímeros o trímeros a través de la coordinación de metales de Cu^{2+} o Zn^{2+}. Resultados similares se obtuvieron con variantes de MBP con 2 histidinas. Tras insertar 2 cisteínas a la lisozima, la misma cristalizó como un trímero donde 3 pares de cisteína coordinaban 4 átomos de Zinc. También se obtuvo otra estructura donde se formo un tetrámero covalentemente unido mediante 4 S-S sin observarse la presencia del metal (236).

3.6 Fusión con proteínas para la cristalización

Las proteínas de fusión han sido muy útiles para mejorar la expresión y plegamiento de proteínas de interés e incluso en algunos casos sirven para asistir en los pasos de purificación. Sin embargo, para estudios cristalográficos la presencia de una proteína de fusión puede añadir heterogeneidad conformacional, dada por la posición relativa entre las dos proteínas que están unidas por una secuencia flexible, dificultando así el proceso de cristalogénesis. En este sentido es que comúnmente se realiza un corte proteolítico para eliminar la proteína de fusión. Sin embargo, fusiones completas han sido cristalizadas al utilizar una secuencia corta y rígida que funcionaba como “linker” entre ambas proteínas. Al analizar las estructuras obtenidas, se vio generalmente que la proteína de fusión aportaba superficies proteicas para la formación de nuevos contactos cristalinos (110). Además, como la estructura de las proteínas de fusión comúnmente utilizadas es conocida, dependiendo del caso, pueden utilizarse como modelos para la determinación de la estructura de la fusión entera mediante reemplazo

molecular (111). Dentro de las proteínas de fusión utilizadas para cristalizar otras proteínas, encontramos a MBP, Trx, GST, GFP y lisozima.

En el primer reporte del uso de fusiones con MBP, se logró determinar la estructura a 2.5 Å del fragmento de la proteína gp21 (residuos 338-425) de la envoltura del virus de HTVL-1. Para esto se fusionó la proteína mediante una secuencia de tan sólo 3 alaninas (237). En otro ejemplo, se logró obtener la estructura del regulador accesorio R (SarR) de *S. aureus* a 2.3 Å fusionado a MBP mediante una secuencia de cinco residuos (AAAEF). Además en este caso algunos residuos cargados de la hélice C-terminal de MBP fueron mutados a alanina (238). Utilizando la misma secuencia espaciadora fue posible obtener las estructuras de las fusiones con MBP del dominio PAZ de Argonauta2 de *Drosophila* (136 AA) a 2.8 Å (239) y el dominio extracelular del receptor de la hormona paratiroidea humana (PTH1R; 174 AA) a 1.95Å (240). Más recientemente, se aplicó la metodología de SER sobre MBP, generándose una serie de 5 variantes diferentes que podrían ser más propensas a cristalizar, para ser utilizadas como proteínas de fusión para la cristalización de otras proteínas blanco. Estas fusiones están conectadas por una secuencia corta de 3 alaninas además de 3 substituciones a alanina en la hélice C-terminal de MBP (108). Mediante esta metodología, se logró obtener las estructuras de 3 proteínas que no habían podido cristalizarse sin la fusión. Este es el caso de la proteína 2-*O*-sulfotransferasa (2OST; 298 AA) de *Gallus gallus* (241), el receptor de la C-quinasa 1 (RACK1A; 324 AA) de *Arabidopsis thaliana* (242) y la proteína Der p 7 (198 AA) de *Dermatophagoides pteronyssinus* (243) que se obtuvieron a 2.65, 2.4 y 2.35Å de resolución respectivamente.

Otra proteína de fusión que ha permitido la cristalización de un blanco que no podía cristalizarse es Trx. En este sentido se fusionó al C-terminal de Trx, el dominio UHM del factor de transcripción Puf60 (100 AA), utilizando 2 secuencias espaciadoras diferentes (GSAM ó GSPPM). Solo con la secuencia GSAM se obtuvieron cristales permitiendo resolver la estructura de la fusión entera a 2.2 Å de resolución mediante reemplazo molecular utilizando la estructura ya conocida de Trx (111).

De manera similar, GFP se utilizó para la cristalización de proteínas pequeñas como ser ubiquitina (8.5 kDa) y un motivo de unión a ubiquitina de la polimerasa de ratón (UBM2; 6 kDa). En este caso los 8 residuos carboxilo terminales de GFP fueron removidos dejándose una secuencia espaciadora de 2 residuos (GS). De esta forma se determinaron las estructuras de GFP-Ubiquitina y GFP-UBM2 a 1.4 y 1.6 Å de resolución (109).

También segmentos proteicos cortos se han fusionado en el N- y C- terminal de lisozima o GST obteniéndose las estructuras de los péptidos de fusión que cristalizaron incluso en condiciones similares a la de la lisozima o GST libres (244).

Finalmente en una aproximación diferente, al reemplazar el tercer bucle interno flexible del receptor β_2-adrenérgico (β_2A-GPCR) por la lisozima T4, se logró estabilizar las hélices 5 y 6 del receptor, permitiendo la cristalización de la fusión y determinar su estructura a 2.4 Å de resolución (245, 246). Recientemente se propuso otra opción a la lisozima T4 como candidato para insertar en loops internos de proteínas de membrana para facilitar su cristalización. Este corresponde a la variante termoestabilizada, apocitocromo b_{562}RIL, optimizada para dicho propósito (247) permitiendo obtener la estructura del receptor de adenosina A_{2A} humano a 1.8 Å de resolución, la más alta obtenida para un receptor tipo GPCR (248).

3.7 Proteínas de unión como chaperonas de cristalización

Las proteínas de unión tienen diversas aplicaciones, incluyendo también su uso como chaperonas de cristalización, es decir, pueden favorecer la cristalogénesis luego de unirse a su blanco. La base de esta estrategia radica en que tras la unión, se pueden estar fijando regiones móviles en la molécula blanco, disminuyendo de esta forma la heterogeneidad conformacional. Adicionalmente, la proteína de unión puede aportar una superficie proteica extra para facilitar los contactos primarios entre moléculas promoviendo la formación de la red cristalina (249). Asimismo, si la estructura del scaffold es conocida, dependiendo del caso, puede ser utilizado como modelo para resolver la estructura del complejo mediante reemplazo molecular (221).

Las primeras moléculas utilizadas como chaperonas de la cristalización correspondieron a los fragmentos Fab o Fv derivados de AcMo. Por ejemplo, en 1995 se logró aislar por primera vez un AcMo contra una proteína de membrana, la citocromo c oxidasa de *Paracoccus denitrificans*. La estructura de esta proteína pudo ser determinada en complejo con un fragmento Fv del AcMo aislado previamente, observándose que el fragmento Fv mediaba la mayoría de los contactos cristalinos (250). Luego de este trabajo, el uso de fragmentos Fab o Fv permitió resolver las estructuras de otras proteínas de membrana, y en algunos casos derivó en estructuras de mejor resolución en comparación con la proteína original sin el fragmento (251). Además, el uso de estos fragmentos también permitió la visualización de conformaciones específicas de algunas proteínas de membrana, como el reciente caso del transportador de leucina bacteriano LeuT. En este caso, se logró capturar al transportador en 2 estados diferentes estabilizados por fragmentos de Ac sensibles al cambio conformacional

(252). También los fragmentos Fab han sido útiles para cristalizar proteínas solubles como fue el caso de la proteína OspA de *Borrelia burgdorferi*, que solo cristalizó como complejo con un Fab (253). Continuando con moléculas derivadas del sistema inmune, se pueden destacar los nanobodies, derivados de Ac de Camélidos. Éstos han sido utilizados para obtener las estructuras de varios complejos. En un ejemplo reciente, nanobodies fueron dirigidos contra una forma truncada de la β2-microglobulina para bloquear la formación de fibras. De esta forma el nanobody prevenía la agregación de la proteína lográndose obtener la estructura del complejo a 2.16 Å, lo que permitió entender los posibles mecanismos de auto-asociación de la proteína (254). A nivel de proteínas de membrana, recientemente un nanobody fue utilizado para bloquear en una única conformación (su estado activado), al receptor adrenérgico humano β(2), facilitando su cristalización y obtención de la estructura (255). Estudios cristalográficos con complejos nanobody-RNasa A, produjeron 6 formas cristalinas diferentes (que difractaban entre 2.5 a 1.1 Å de resolución). Además, se introdujeron 2 metioninas adicionales en el nanobody mediante "shotgun Met scanning" (llegando a un total de 5 metioninas), para ser sustituidas por seleno-metioninas (SeMet). De esta forma se determinó que con esta modificación el nanobody con 5 SeMet podía proveer de suficiente poder de faseo para resolver la estructura del complejo mediante dispersión anómala única (SAD), sin la necesidad de incorporar SeMet en la proteína blanco (256). Recientemente se ha desarrollado un protocolo general para la generación de nanobodies para la cristalización de proteínas blanco, que incluye desde los pasos de inmunización de Llamas hasta la selección de nanobodies de alta afinidad por "phage display" (257).

Otro scaffold que ha servido para la obtención de estructuras cristalográficas son las DARPins. De esta manera se obtuvieron las estructuras de complejos de DARPins con: la quinasa bacterial aminoglicosido fosfotransferasa (APH) (258), la quinasa eucariota Plk1 (259), la caspasa-2 humana (260), la proteína RBP perteneciente al fago lacctococal TP901-1 (261), y la proteína de membrana de *E. coli* AcrB (262). Más recientemente se lograron obtener las estructuras de DARPins que unían a la proteína quinasa activada por mitógeno, ERK2 en su estado fosforilado y desfosforilado. Se vio además que las DARPins unían esencialmente la misma región, pero eran capaces de reconocer el cambio conformacional en el loop de activación tras la fosforilación de ERK2 (263). Esto muestra que las proteínas de unión no solo fijan regiones móviles y/o aportan superficies proteicas para la generación de contactos cristalinos, sino que también permiten obtener las estructuras en estados conformacionales que pueden ser biológicamente importantes, ayudando a comprender distintos procesos biológicos.

Recientemente, se propuso una aproximación diferente en donde se inserta en la proteína blanco un motivo para la interacción con otra proteína y se logra cristalizar así el complejo. Es así que el grupo de Geoffrey Waldo, propuso la inserción de una horquilla compuesta por las hebras-β 10 y 11 de GFP (39 residuos) en un loop de superficie de la proteína blanco. De esta forma la proteína adquiere la capacidad de interaccionar y unirse a la variante 1-9 de GFP pudiéndose cristalizar el complejo (264). Como prueba de concepto se insertaron las hebras 10 y 11 de GFP en un loop de la proteína fluorescente sfCherry. La proteína conteniendo dicha horquilla pudo acomplejarse con el fragmento 1-9 de GFP, dando lugar al desarrollo de fluorescencia tras la interacción y el complejo pudo cristalizarse obteniéndose la estructura a 2.6 Å de resolución (264). Esta estrategia puede ser útil por ejemplo para cristalizar proteínas de membrana donde en lugar de insertar la lisozima T4 en un loop, puede insertarse esta secuencia para formar posteriormente el complejo con GFP1-9. De esta manera GFP puede realizar contactos cristalinos adicionales favoreciendo la formación del cristal y la obtención de la estructura.

Otra aproximación ha sido desarrolla, pero para compuestos químicos pequeños, en donde se generan pequeños cristales de un complejo poroso capaz de unir la molécula blanco. Una vez en el poro del cristal, las moléculas quedan retenidas y orientadas de manera regular mediante interacciones con la red que forma el cristal, pudiéndose de esta manera determinar la estructura del complejo sin la necesidad de cristalizar el compuesto aislado (265).

Si bien la obtención de cristales que difracten a alta resolución, es uno de los obstáculos más importantes que enfrenta la biología estructural, diversas estrategias se han desarrollado para aumentar las chances de tener éxito en la obtención de la estructura de la proteína blanco. Además, estas estrategias no son excluyentes sino que pueden ser complementarias, por ejemplo el uso de la proteína MBP donde se adicionaron mutaciones de superficie para reducir la entropía configuracional y favorecer así la cristalización (108).

OBJETIVOS

Como se describió en la parte introductoria de esta tesis, varios obstáculos en el área de la expresión de proteínas recombinantes y cristalografía deben ser sorteados antes de llegar a la obtención de un cristal que genere datos de difracción útiles. Lograr la información a nivel molecular y estructural que nos permita entender la función de la proteína de interés así como las interacciones de esta proteína con otras moléculas es dependiente en gran medida de la superación de dos problemas técnicos principales: a) producir suficientes cantidades de la proteína recombinante en forma soluble y funcional, b) obtener cristales con poder de difracción.

Este trabajo de Tesis intentó abordar estos problemas proponiendo nuevas metodologías para la producción de proteínas recombinantes y nuevas aproximaciones experimentales para la obtención de cristales de alta calidad. El desarrollo de estas tecnologías nos ha permitido también abordar el desafío de generar nuevas moléculas de unión que no solamente podrían ayudar al proceso cristalográfico sino también actuar como moléculas capaces de unirse a sitios específicos de proteínas blanco que puedan tener un interés en el área biomédica.

Teniendo todo esto en cuenta es que en el presente trabajo de Doctorado, nos centramos en aportar nuevas herramientas para facilitar la evaluación de la expresión soluble de PRs, trabajar en la generación de proteínas de unión con propiedades favorables para la inhibición enzimática y finalmente, desarrollar nuevas herramientas o estrategias para la cristalización de proteínas.

Objetivo general:

Diseño e implementación de nuevas tecnologías aplicadas a la mejora de la solubilidad, unión y cristalización de PRs.

Objetivos específicos:

1. Generación de nuevas herramientas para la expresión soluble de PRs.
    a. Generación de una serie de vectores para facilitar la evaluación de condiciones de expresión de PRs.
    b. Uso de una nueva proteína de fusión (CelDnc) como potenciador de la solubilidad de PRs.

2. Aplicación y caracterización de una nueva librería de Afitinas como inhibidores enzimáticos de glicosidasas.
    a. Generación de proteínas de unión contra la endogluconasa CelD utilizando 2 tipos diferentes de librerías de afitinas.
    b. Determinación de la afinidad de las afitinas obtenidas para CelD y lisozima mediante ITC y Biacore y de su estabilidad térmica mediante DSC.
    c. Determinación de la inhibición enzimática de las afitinas anti-CelD y anti-lisozima.
    d. Obtención y análisis de las estructuras cristalográficas de los complejos afitina-CelD y afitina-lisozima.
3. Generación de nuevas herramientas para la cristalogénesis de PRs.
    a. Puesta a punto del uso de la proteína CelD como proteína de fusión para la cristalización de PRs.
    b. Generación de superficies de unión en CelD para obtener cristales de complejos entre CelD y la proteína blanco.

RESULTADOS

1. Generación de nuevas herramientas para la expresión soluble de PR.

"Generation of a vector suite for protein solubility screening"

Agustín Correa[1], Claudia Ortega[1], Gonzalo Obal[2] , Pedro Alzari[3] , RenaudVincentelli[4] and Pablo Oppezzo[1]*

1 Recombinant Protein Unit, Institut Pasteur de Montevideo, Montevideo, Uruguay
2 Protein Biophysics Unit, Institut Pasteur de Montevideo, Montevideo, Uruguay
3 Unité de Microbiologie Structurale, Institut Pasteur, Paris, France
4 Centre National de la Recherche Scientifique, Aix-Marseille Université, CNRS UMR7257, AFMB, Marseille, France

Frontiers in Microbiology, February 2014, Vol. 5, 1-9.

Las proteínas recombinantes han adquirido un rol muy importante, siendo utilizadas tanto en áreas básicas como aplicadas. Sin embargo, muchas veces para lograr una expresión soluble y homogénea de la proteína blanco es necesario evaluar diferentes condiciones de expresión como ser diferentes condiciones de cultivo, distintas proteínas de fusión, diversos promotores, entre otras. Esto lleva a que el gen de interés tenga que ser clonado en diferentes vectores. Nuevos métodos de clonado, independientes del uso de enzimas de restricción permiten la inserción del gen de interés en distintos vectores de expresión en simultáneo, permitiendo en consecuencia la evaluación de diversas condiciones de expresión.

Por otro lado, se ha visto que la utilización de proteínas de fusión puede aumentar significativamente la expresión soluble de la proteína blanco (91). Al día de hoy existen muchas proteínas que se han utilizado para la expresión de diferentes blancos como MBP, SUMO, Trx y DsbC entre otras. Debido a que aún no se ha encontrado una proteína de fusión genérica, que funcione para todos los blancos posibles, es que es necesario evaluar diferentes fusiones. Por lo tanto un problema común a la hora de expresar una proteína recombinante es la necesidad de evaluar diferentes proteínas de fusión, lo que lleva a requerir de un simple y eficiente método de clonado.

En esta primera parte de la Tesis, utilizando la metodología de clonado RF generamos una serie de 12 vectores para evaluar las condiciones de expresión de un gen de interés, que permite analizar en simultaneo 2 promotores y 5 proteínas de fusión diferentes o la proteína blanco sin fusión. Debido a que el sitio de inserción es el mismo en todos los vectores, el mismo inserto o producto de PCR puede introducirse en simultáneo en la serie completa. Además exploramos el uso de una nueva proteína de fusión, una versión truncada de la endogluconasa CelD termostable (CelDnc), de manera de ampliar el repertorio de proteínas de fusión disponibles y por ende las probabilidades de tener éxito en la expresión soluble de la proteína recombinante. Esta proteína se expresa de forma soluble y homogénea en cantidades masivas en *E. coli*, es termostable (Tm: 74˚C) y permanece activa incluso luego de la presencia de 8M urea (266). Debido a estas características la forma truncada de CelD fue incluida como proteína de fusión para determinar si alguna de estas propiedades pueden ser transmitidas a la proteína blanco mejorando así su solubilidad/estabilidad.

La serie fue generada en base a dos vectores comerciales, pQE80 (promotor T5) y pET32a (promotor T7) donde se insertaron como proteínas de fusión: MBP, SUMO, Trx, DsbC y CelDnc, o la posibilidad sin proteína de fusión. Los vectores codifican para un sitio de corte para la endoproteasa TEV, un HisTag amino terminal y un strepTag II carboxilo terminal. En una

primera instancia se logró clonar con alta eficiencia en la serie completa los genes para la proteína DprE1 (51 kDa) de *Mycobacterium smegmatis*, la proteína fluorescente GFP (26.9 kDa) de *Aequorea victoria*, y la quinasa de *Leishmania major* MPK4 (41.5 kDa).

Con los vectores conteniendo los genes de GFP y DprE1, se realizó una evaluación de la expresión y validación de los vectores utilizando 2 temperaturas de inducción (37 y 17˚C) en medio TB (Terrific Broth) y utilizando la cepa de *E. coli* Rosetta-pLysS. Luego de la expresión, purificación por HisTag y corte con TEV de las 48 condiciones en forma manual, se evaluaron los niveles de expresión mediante SDS-PAGE 96x (E-PAGE 96, Invitrogen). Para el caso de GFP, todas las construcciones mostraron expresión de las diferentes fusiones, y también un correcto corte con la proteasa TEV, demostrándose el correcto funcionamiento de los 12 vectores generados. Al analizar los resultados obtenidos para DprE1, se observó que sin fusión no se obtenía expresión de la proteína, sin embargo buenos rendimientos fueron alcanzados para las fusiones con MBP, Sumo, Trx y CelDnc. Además un correcto corte y liberación de la proteína DprE1 tras la incubación con TEV fueron obtenidos para estas construcciones. Interesantemente, con las fusiones con CelDnc se obtuvieron rendimientos iguales o incluso mayores que los obtenidos para las proteínas de fusión ya conocidas (MBP, SUMO, Trx y DsbC), mostrando el potencial de esta nueva proteína de fusión como potenciador de la solubilidad de proteínas blanco.

Para determinar si la proteína DprE1 permanece soluble y homogénea una vez cortada la fusión, es que se utilizó la construcción pT5-CelD-DprE1 para realizar una expresión a mayor escala (1 lt). De esta manera se pudo constatar que DprE1 se mantenía soluble luego del corte con TEV y en estado monomérico tras la purificación por gel filtración obteniéndose un rendimiento final de 7 mg/lt. Más aún, DprE1 mantenía unión al FAD (Flavín Adenín Dinucleótido), por lo que estaría bien plegada gracias a la fusión con CelDnc.

Con la proteína MPK4 también se realizó una expresión, purificación y análisis, pero de forma automática en donde sólo se evaluó sólo una temperatura de inducción, 17˚C. La cepa y medio de cultivo fueron los mismos que para GFP y DprE1. En este caso, la única condición en donde se visualizó una banda correspondiente a MPK4 tras el corte con TEV fue para la construcción con la fusión con DsbC. Tratando de obtener una mayor cantidad de la proteína de interés y validar los resultados, se realizó un escalado de esta condición. Los resultados mostraron que tras el corte con TEV, la gran mayoría de la proteína precipitaba. Al analizar la fusión entera por gel filtración, se encontró que la misma presentaba un estado decamérico, lo que podría ser utilizado para ensayos de cristalogénesis tras optimizar las etapas de purificación.

Todos estos resultados permitieron proponer, una herramienta que permite el fácil clonado del gen de interés en 12 vectores en simultáneo, para el análisis de expresión tanto en forma manual como automática. Además, este es el primer reporte donde se muestran las propiedades solubilizadoras de la proteína CelDnc y por ende su uso como proteína de fusión, pudiendo ser útil para la expresión de otras proteínas blanco.

Los resultados obtenidos en este primer capítulo de la Tesis de Doctorado fueron recientemente publicados en la revista **Frontiers in Microbiology**.

frontiers in
MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 25 February 2014
doi: 10.3389/fmicb.2014.00067

Generation of a vector suite for protein solubility screening

Agustín Correa[1], Claudia Ortega[1], Gonzalo Obal[2], Pedro Alzari[3], Renaud Vincentelli[4] and Pablo Oppezzo[1]*

[1] Recombinant Protein Unit, Institut Pasteur de Montevideo, Montevideo, Uruguay
[2] Protein Biophysics Unit, Institut Pasteur de Montevideo, Montevideo, Uruguay
[3] Unité de Microbiologie Structurale, Institut Pasteur, Paris, France
[4] Centre National de la Recherche Scientifique, Aix-Marseille Université, CNRS UMR7257, AFMB, Marseille, France

Edited by:
Eduardo A. Ceccarelli, Universidad Nacional de Rosario, Argentina

Reviewed by:
Jun-Jie Zhang, Chinese Academy of Sciences, China
Grzegorz Wegrzyn, University of Gdansk, Poland

***Correspondence:**
Pablo Oppezzo, Recombinant Protein Unit, Institut Pasteur de Montevideo, Mataojo 2020, Montevideo 11400, Uruguay
e-mail: poppezzo@pasteur.edu.uy

Recombinant protein expression has become an invaluable tool for academic and biotechnological projects. With the use of high-throughput screening technologies for soluble protein production, uncountable target proteins have been produced in a soluble and homogeneous state enabling the realization of further studies. Evaluation of hundreds conditions requires the use of high-throughput cloning and screening methods. Here we describe a new versatile vector suite dedicated to the expression improvement of recombinant proteins (RP) with solubility problems. This vector suite allows the parallel cloning of the same PCR product into the 12 different expression vectors evaluating protein expression under different promoter strength, different fusion tags as well as different solubility enhancer proteins. Additionally, we propose the use of a new fusion protein which appears to be a useful solubility enhancer. Above all we propose in this work an economic and useful vector suite to fast track the solubility of different RP. We also propose a new solubility enhancer protein that can be included in the evaluation of the expression of RP that are insoluble in classical expression conditions.

Keywords: recombinant proteins, solubility, expression, vector, cloning, high-throughput

INTRODUCTION

Recombinant protein production has become a routine practice in many laboratories from academic to industrial fields. Several hosts are available for protein production among them, *Escherichia coli* has been by far the most widely used. Some advantages of this host is the low cost, infrastructure of implementation, easy handling, high yield production, and an ever increasing set of tools and genetic information useful for the expression of challenging targets. Despite its importance and utility, recombinant proteins (RP) not always are produced in a soluble and homogeneous state. For these "difficult to express" proteins, several approaches have been developed in order to overcome the problems associated with insolubility. Some parameters that can affect protein expression are: induction temperature, promoter strength, use of specific *E. coli* strains, co-expression of molecular chaperones or biological partners and the use of different solubility enhancer or fusion proteins (Correa and Oppezzo, 2011). In the last decade, the advent of high-throughput screening methods have facilitated the evaluation of hundreds of conditions generated from the combination of the mentioned parameters in order to find one that gives a soluble protein (Vincentelli et al., 2011; Vincentelli and Romier, 2013). However, to exploit all these variables it is necessary to have a method for cloning the target gene in many different vectors in a fast and simple manner. Several techniques were recently generated to facilitate the cloning of target genes in a parallel way, in which the same insert can be introduced into different expression vectors simultaneously. Among these methods are the Gateway technology [Invitrogen, (Esposito et al., 2009)], In-Fusion technology, [Clontech, (Berrow et al., 2007)], Ligase Independent Cloning, (Aslanidis and de Jong, 1990), and Restriction Free Cloning, [RF cloning, (Unger et al., 2010)]. With these methodologies, the use of restriction endonucleases is avoided, so no special sequence requirements are necessary enabling the development of high-throughput technologies for molecular cloning (Cabrita et al., 2006; Berrow et al., 2007; Curiel et al., 2010; Unger et al., 2010; Luna-Vargas et al., 2011).

In this work, we have modified two commonly used commercial vectors (pET32a and pQE80L, T7 and T5 promoters respectively) for *E. coli* protein expression. We generated 12 different vectors introducing the same sequence at the insertion site, and important features for protein purification like N-terminal $(His)_6$ tag (Murphy and Doyle, 2005), TEV cleavage site, and C-terminal StrepTag II (Schmidt and Skerra, 2007), in order to set up a high-throughput cloning and purification protocol. The cloning strategy used for the development of the vectors as well as for cloning the target genes on the entire suite is based in the "RF cloning methodology" (Unger et al., 2010). The data reported here, describe the application of an easy methodology to clone any target in 12 different vectors with only two primers. In order to evaluate and find a condition for soluble protein expression, different promoters and solubility enhancer fusion proteins were included in these vectors. Concerning protein solubility enhancers, the target gene can be fused as a C-terminal partner with maltose binding protein (MBP; Kapust and Waugh, 1999), thioredoxin A (Trx; LaVallie et al., 2000), small ubiquitin-like modifier protein (SUMO; Marblestone et al., 2006), disulfide bond isomerase C (DsbC; Nozach et al., 2013), and Histag alone in a T5 or T7 promoter context.

Finally, we propose a new fusion protein which appears to be an efficient solubility enhancer for the RP with previous solubility problems and is included in the vector suite. This solubility enhancer corresponds to a truncated construct of the endoglucanase CelD (CelDnc) from *Clostridium thermocellum*. This is a thermostable protein, highly expressed in *E. coli* system and more interestingly, this molecule maintains a full activity even in the presence of 8M Urea implying a very high stability of its native structure (Chaffotte et al., 1992). All these characteristics make CelDnc a good candidate to study the solubility enhancing properties when fused a target protein. As a proof of concept, we fused to CelDnc the decaprenylphosphoryl-β-D-ribofuranose-2′-epimerase (DprE1) protein from *Micobacterium smegmatis* (Neres et al., 2012) a difficult protein to express in *E. coli* (<0.4 mg/l) and we successfully improved this expression obtaining high yields of soluble and functional monomeric protein.

In summary, here we illustrate how to generate in any laboratory an economic and useful vector suite to fast track the solubility of different RP targets and we propose a new solubility enhancer protein that can be included in the evaluation of the expression of RP that are insoluble in classical expression conditions.

RESULTS

CONSTRUCTION OF A NEW VECTOR SUITE

Aiming to achieve a fast and economical way to evaluate the solubility of RP, we selected two commonly used expression vectors pQE-80L (Qiagen) and pET-32a (Novagen) as the starter plasmids for the suite generation thus giving rise to T5 or T7 based vectors. In order to provide a parallel cloning of the target gene and an easy protein purification method, all the generated vectors contain the same insertion site and antibiotic resistance (ampicillin), an N-terminus His-Tag with the tobacco etch virus (TEV) recognition site and a C-terminus strep-Tag II (**Figure 1**; **Table 1**). In addition, we introduced several solubility enhancing proteins including MBP, Trx, DsbC, SUMO, and CelDnc, in combination with the two promoters (T5 or T7). An extra serine residue was added after the TEV site to decrease steric effects and improve cleavage. This can be avoided by not including it in the forward primer. This extra codon also generates a *Bam*HI site at the beginning of the gene so it can be useful for analysis of clones or to do a restriction based method if preferred (**Figure 1**).

VALIDATION OF THE NEW VECTOR SUITE

In order to evaluate the expression capabilities and functionality of this new vector suite we selected green fluorescent protein (GFP) as control protein and two "difficult to express" RP such as DprE1 and the MAP kinase 4 from *Leishmania major* (MPK4). All of them were cloned into 12 different vectors and their expression was evaluated. The results showed that all the GFP constructs were produced soluble and at the expected molecular weight. Fractions treated with TEV showed the correct cleavage and release of GFP protein and fusion partner (**Figure 2A**). The construct DsbC-GFP under the control of T7 promoter was the less productive when working at 37°C. This was over-passed when the expression was done at 17°C over night (ON) where an increment of cleaved proteins was obtained in most of the cases (**Figure 2A**).

For the case of DprE1 constructs, we can see that despite a correct growth and induction conditions in the culture, it was not possible to obtain any expression of this RP when fused only to a Histag. In contrast, fusion of DprE1 with MBP, Sumo, Trx, and CelDnc give a good soluble production and only low yields account for the DsbC/DprE1 construct (**Figure 2B**; **Table 2**). Also, there was an effect of the induction temperature and promoter strength in protein expression where DprE1 was expressed with higher yields at 37°C compared to 17°C and with the T5 promoter compared with T7 for most of the cases. Interestingly, our results suggested that DprE1 fused with CelDnc (in the condition T5-37°C) appear to be one of the most overexpressed fused proteins. For the case of DprE1/CelDnc in T7 at 17°C, there was no cell growth. Finally, the treatment with TEV revealed that DprE1

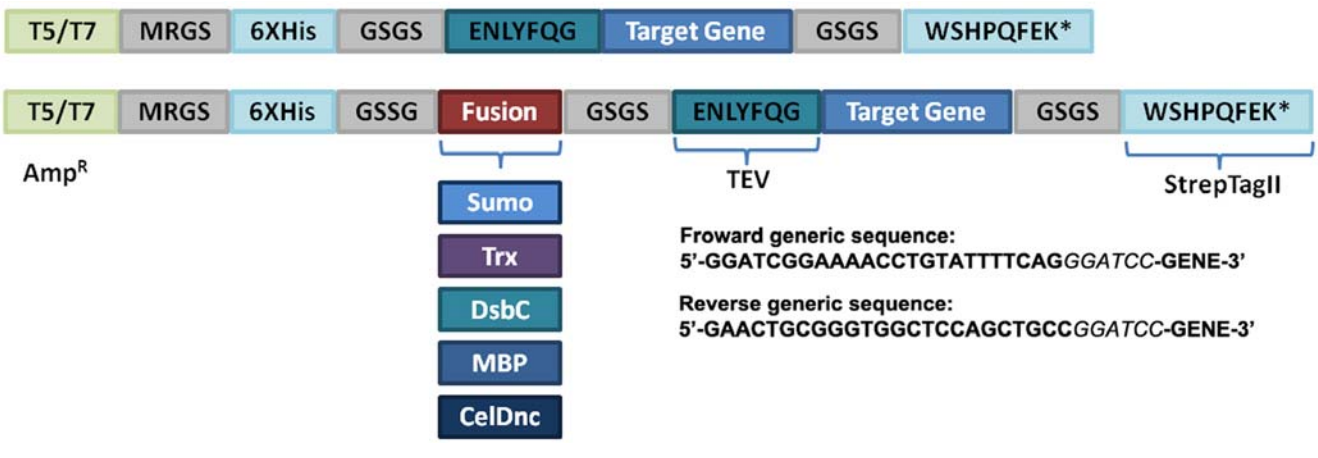


FIGURE 1 | Schematic representation of vector suite. The generic module with the different characteristics is shown. T5/T7, promoter type; Amp^R, ampicillin resistance; tobacco etch virus protease recognition site (TEV), 6xHis, His-Tag; Sumo, Trx, DsbC, MBP, and CelDnc are the different fusion proteins. The generic sequences to add to the forward and reverse primers are indicated, italic letters correspond to the *Bam*HI restriction site. * represent a stop codon.

Table 1 | Primer list for vector generation, cloning, and sequencing.

Primer	Sequence 5′–3′	Characteristics
CelDwtNFor	GGATCGGAAAACCTGTATTTTCAGGGATCCATGACCATGATTACGAATTCCCGG	Cloning of CelDwt
CelDwtCRev	GCTGCAGGTCGACGCCAAGATCCTTTTTTATATTGGTAATTTCTCGATTACCCT	
CelDtruncNFor	GGATCGGAAAACCTGTATTTTCAGGGATCCTCGGGATTGATTGAGACCAAAGTG	Cloning of CelDnc
CelDtruncCRev	GCTGCAGGTCGACGCCAAGATCCTTTTTTAAGCAGAATTATAGTTGACAAATCCGG	
QE3790For	GGATCGGAAAACCTGTATTTTCAGGGATCCATGGGCGCGGTACCCTCACTGACG	Cloning of Rv3790
QE3790Rev	GCTGCAGGTCGACGCCAAGATCCTTTTTTAGAGCAGTTGCAGGCGCCTGGCCATG	
CelDInsFor	ATGAGAGGATCGCATCACCATCACCATCACGGATCTTCGGGATTGATTGAGACCAAAGTGTC	Insertion of CelDnc as a fusion partner
CelDInsRev	GGATCCCTGAAAATACAGGTTTTCCGATCCGCTACCAGCAGAATTATAGTTGACAAATC	
strepCterFor	TCCGACATGGCCAGGCGCCTGCAACTGCTCGGATCCGGCAGCTGGAGCCACCCGCAGTTC	Insterion of a C-terminus strepTagII
strepCterRev	TGGCTGCAGGTCGACGCCAAGATCCTTTTTTACTTTTCGAACTGCGGGTGGCTCCAGCTG	
SumoFor	CATCACCATCACCATCACGGATCTTCGGGAATGTCGGACTCAGAAGTCAATCAAG	Insertion of Sumo as a fusion partner
SumoRev	CTGAAAATACAGGTTTTCCGATCCGCTACCATACGTAGCACCACCAATCTGTTC	
TrxFor	CATCACCATCACCATCACGGATCTTCGGGAATGAGCGATAAAATTATTCACCTG	Insertion of TrxA as a fusion partner
TrxRev	CTGAAAATACAGGTTTTCCGATCCGCTACCGGCCAGGTTAGCGTCGAGGAACTC	
MBPFor	CATCACCATCACCATCACGGATCTTCGGGAATGAAAACTGAAGAAGGTAAACTG	Insertion of MBP as a fusion partner
MBPRev	CTGAAAATACAGGTTTTCCGATCCGCTACCATTAGTCTGCGCGTCTTTCAGGGC	
DsbCFor	CATCACCATCACCATCACGGATCTTCGGGAGATGACGCGGCAATTCAACAAACGTTAGCC	Insertion of DsbC as a fusion partner
DsbCRev	CTGAAAATACAGGTTTTCCGATCCGCTACCTTTACCGCTGGTCATTTTTTGGTGTTCGTC	
T5T7For	TTTGTTTAACTTTAAGAAGGAGATATACATATGAGAGGATCGCATCACCATCACCATCAC	Transfer of the entire cassette to pET32a vector
T5T7Rev	CAGTGGTGGTGGTGGTGGTGCTCGAGTGCGGCTTGGCTGCAGGTCGACGC	
GFPFor	GGATCGGAAAACCTGTATTTTCAGGGATCCAGCAAAGGAGGAGAACTTTTC	GFP cloning
GFPRev	GAACTGCGGGTGGCTCCAGCTGCCGGATCCTCAAAGCTTTTTGTAGAGCTCATC	
MPK4For	GGATCGGAAAACCTGTATTTTCAGGGATCCATGGCTCAACTCGTCCCTTTAGC	MPK4 cloning
MPK4Rev	GAACTGCGGGTGGCTCCAGCTGCCGGATCCCTATTCGTTCAATTGTGAATGGG	
MBPSeqFor	CGCTGGCGCGAAAGCGGGTC	MBP sequencing
CelDSeqFor	GTGCCCTGGAGCAGTGCCGC	CelD sequencing

RP fused with all these enhancers remains in a soluble state confirming an important improvement expression after usage of this vector suite (**Figure 2B**).

Concerning MPK4 our results showed that of the 12 constructs only 2 gave a band at the expected molecular weight. These correspond to the construct pT7-DsbC-MPK4 and pT7-MBP-MPK4 (**Figure 2C**). In both cases TEV protease was able to cleave the fusion but only in the pT7-DsbC-MPK4 constructs it was possible to get a soluble protein after cleavage (**Figure 2C**, TEV treatment section). In order to confirm this result and validate our suite vector we proceed to perform a large scale purification with this construct. Our results showed that after protein purification by IMAC it is possible to obtain the DsbC-MPK4 fusion in a soluble manner and with a yield of 6 mg/l (**Figure 2D**). Oligomeric state analysis of the DsbC-MPK4 fusion, revealed that the eluted peak is maintained as a soluble decameric oligomer with an apparent molecular weight of around 650 kDa (**Figure 2D**). This result was verified by dynamic light scattering (data not shown). Despite the fact, a great part of the MPK4 protein precipitate after TEV treatment, an interesting and scalable amount of this protein remains in a soluble form (**Figure 2D**).

Altogether these results, underline the importance of this new vector suite as an improved tool for the soluble expression of DprE1 and MPK4 proteins and suggest that it can be very valuable for the expression of other "difficult to express" RP.

USE OF ENDOGLUCONASE D VARIANT (CelDnc) FOR THE SOLUBLE EXPRESSION OF DprE1 PROTEIN

After expressing the new construct (CelDnc), we found out that it is expressed at high yields (>400 mg/l) in a soluble monomeric and functional form which in turn maintains thermostable characteristics as the entire version (**Figures 3A,B**). So, we wondered if this extreme solubility and stability could help in the production and folding of other target proteins. In this regard, we fused CelDnc to the N-terminus of the protein DprE1. The results showed that the fusion was successfully produced in a soluble manner and that after TEV treatment and gel filtration purification it remains soluble, monomeric and it was able to retain a FAD binding property

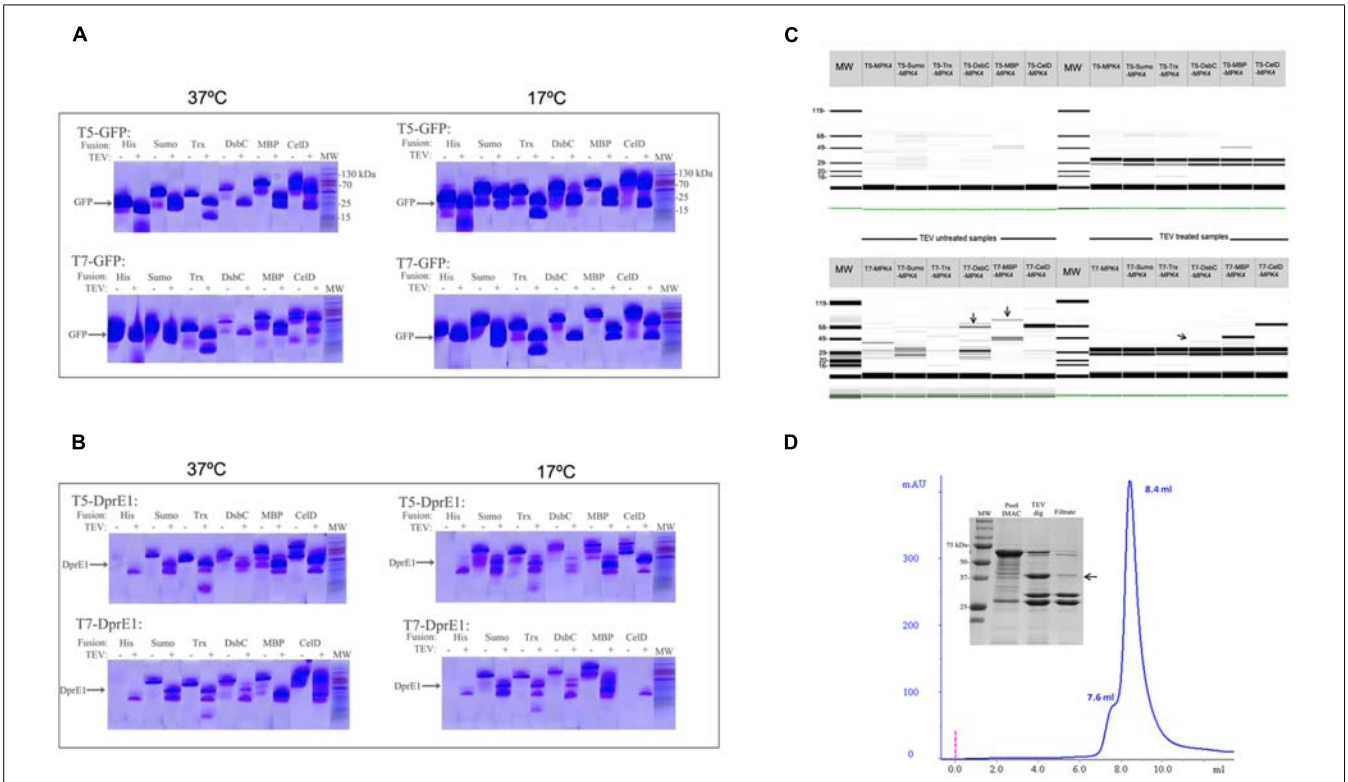


FIGURE 2 | Protein production screening in the vector suite. Panels **(A,B)** corresponds to the E-PAGE 96 acrylamide gels for the expression screening of GFP and DprE1, respectively. The incubation with TEV protease for fusion cleavage is indicated with a +sign over the corresponding lines. Cleaved target protein at the expected molecular weight (MW) is depicted. Additionally, induction temperatures are indicated over each panel. **(C)** Expression screening for MPK4 at 17°C using a Labchip GX II (Caliper, USA) microfluidic detection system. Arrows indicate the presence of a band with the expected molecular weight. Construct names are provided over each gel line. Solubility improvement with vector suite is indicated by arrows. **(D)** Analytical size exclusion chromatography (SEC) of the IMAC purified fraction of DsbC-MPK4. Peaks at 7.6 and 8.4 ml correspond to the exclusion volume and the 600 kDa decameric form of DsbC-MPK4, respectively. The 12% SDS-PAGE shows the fusion protein obtained by IMAC purification and DsbC-MPK4 digested by TEV protease. The expected molecular weight of MPK4 (41.7 kDa) is indicated by an arrow.

Table 2 | Expression screening of DprE1 protein.

Construct name	Fusion protein	MW DprE1 fusions (kDa)	Yield at 37°C (mg/l)	Yield at 17°C (mg/l)
T5 promoter				
pT5-DprE1	Only HisTag	53.7	0.4	0.2
pT5-Sumo-DprE1	Sumo	65.5	12.3	14.1
pT5-Trx-DprE1	Trx	65.8	14.8	10.4
pT5-DsbC-DprE1	DsbC	77.4	6.2	4.7
pT5-MBP-DprE1	MBP	94.3	15.4	11.3
pT5-CelD-DprE1	CelDnc	114.8	19.5	12.8
T7 promoter				
pT7-DprE1	Only HisTag	53.7	0.1	0.2
pT7-Sumo-DprE1	Sumo	65.5	11.6	10.1
pT7-Trx-DprE1	Trx	65.8	12.4	9.8
pT7-DsbC-DprE1	DsbC	77.4	8.1	3.9
pT7-MBP-DprE1	MBP	94.3	12.8	15.8
pT7-CelD-DprE1	CelDnc	114.8	19.2	ND

After purification by IMAC, concentration of the entire fusions and yield was determined at 280 nm taking into account the different extinction coefficients. The expected molecular weight as well as construct name and characteristics are indicated.

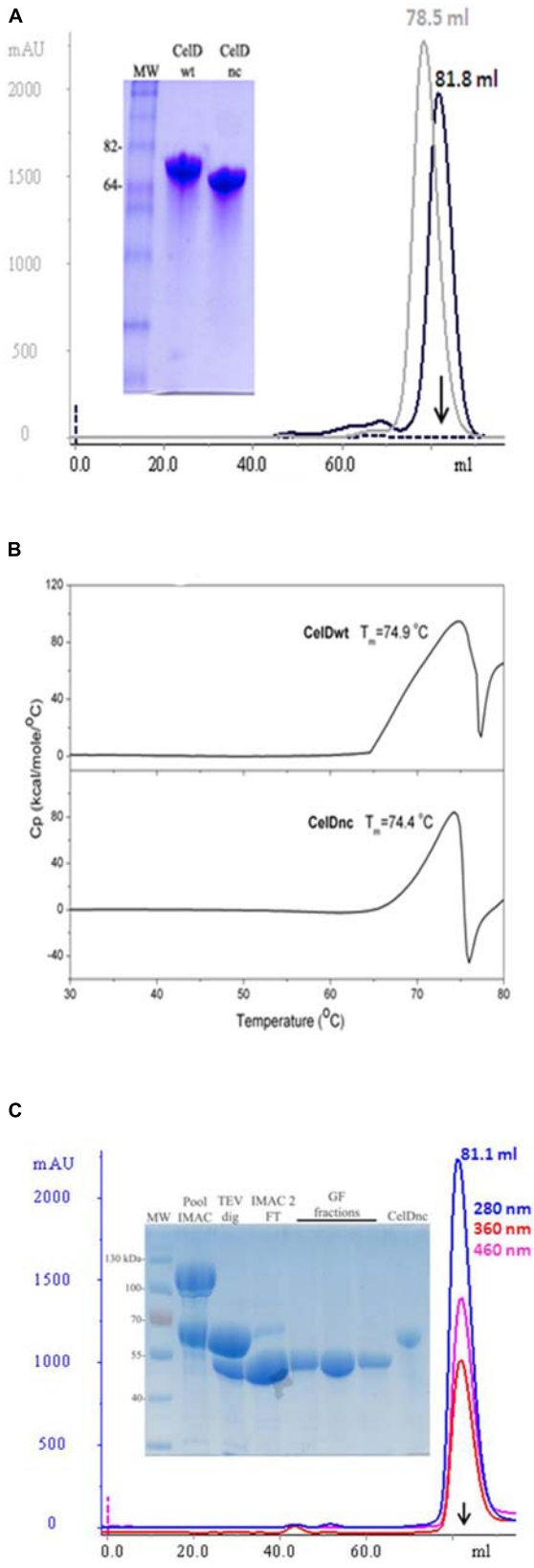


FIGURE 3 | Continued

FIGURE 3 | Continued
(A) Analysis of the purity and monomeric states of CelDwt (gray) and CelDnc (black). SEC was performed in a Superdex 200 16/60 and protein purity evaluated in a 10% SDS-PAGE. **(B)** Differential scanning calorimetry (DSC) curves of CelDwt (top panel) and CelDnc (bottom panel). Determined melting temperature (T_m) is indicated for each case. **(C)** Large scale expression and purification of DprE1. DprE1 was fused to CelD, expressed, and purified by IMAC. After TEV cleavage and second IMAC purification, the monomeric state was confirmed by SEC in a Superdex 200 16/60. FAD binding properties of DprE1 are confirmed by peaks at 360 nm (red) and 460 nm (pink). Purity of DprE1 (53.7 kDa) was evaluated by 12% SDS-PAGE. CelDnc (61.1 kDa) was added as a control. Arrows indicates the retention volume for BSA (66.5 kDa).

as expected for this protein (peaks at 360 and 460 nm; **Figure 3C**). The final yield was of 7 mg/l which corresponds to more than 17 times improvement in soluble protein expression when compared with no fusion (<0.4 mg/l). Moreover, the same experiment done with MBP fusion resulted in a final yield for DprE1 of 2.8 mg/l (data not shown), demonstrating the usefulness of CelDnc as a solubility enhancer of RP.

These results suggest that the construct CelDnc is an interesting new solubility enhancer that could be taken into account for the expression screening of "difficult to express" RP.

DISCUSSION

Purified and soluble proteins are essential tools in academic, industrial and medical areas. The knowledge of the molecular structure of individual proteins allow addressing important questions about the physiological function of these molecules, so as to know the biochemical and regulatory pathways in which they are implicated. However, a common scenario is that the first attempt for obtaining soluble protein often fails, requiring the optimization of many parameters increasing production costs and time. One of the standard procedures to circumvent this problem is to screen a series of constructs to identify the optimal vector and culture conditions able to produce enough soluble protein. This may also include the expression of the full-length protein, mutated and/or truncated variants, as well as specific domains of RP (Dahlroth et al., 2006; Yumerefendi et al., 2010). Series of fusion partners may also be investigated for their effects on driving enhanced expression or their capacity to capture and purify the target protein quickly with minimal impurities (Young et al., 2012).

In this work, we describe the generation of a vector suite composed of 12 different expression vectors using the RF cloning method. This suite engages the expression of the RP with strong promoters such as T7 or T5, with N-terminus His-tag, a TEV specific cleavage site and a C-terminus StrepTag II as well as different fusion proteins such as Sumo, Trx, DsbC, MBP, and CelDnc. All these vectors contain the same site of insertion in order to enable a parallel cloning for solubility screening and the posterior large scale purification in a simple and general manner (IMAC purification, TEV cleavage and dialysis, 2nd IMAC). The suite is based on the commonly used pET and pQE vectors and presents no major changes in expression or sequencing protocols. The cloning strategy occurs in an insert-sequence independent manner, with the additional advantage that no restriction site or extra aminoacids are added to the N-terminus of the expressed protein after TEV cleavage, apart from the last glycine residue. As

purification features we selected the use of the HisTag, because it has demonstrated to be very versatile, cheap and to work well in small and large scale purifications (Schafer et al., 2002; Steen et al., 2006). Additionally, if the stop codon of the target gene is omitted, an additional purification tag, the strepTag II is expressed in the C-terminus of the target protein. This last can be useful if degradation intermediates appear by coupling IMAC purification with StrepTacting purification only a product with an intact N- and C-terminus will be purified. Also the purification via the StrepTag II showed to be very useful for proteins that are expressed in low abundance where usually purification by IMAC gives many contaminants from the host (Magnusdottir et al., 2009). Finally the TEV site was chosen for protein cleavage as it has demonstrated to be very specific, work well at low temperatures and can be produced in the laboratory with high yield reducing production costs (van den Berg et al., 2006). Moreover, it was shown that the last residue of the cleavage site (Gly) can be changed for all the other residues except for proline for an expense in cleavage efficiency, so if a protein with a native N terminus is needed it can be taken into account (Kapust et al., 2002).

The suite was tested with GFP, and we found out that in all cases there were expression and cleavage with TEV demonstrating that all the vectors worked well. By using this suite of vectors the high-throughput screening for soluble expression could be easily achieved manually or automatically as it was demonstrated for the expression of GFP, DprE1 and MPK4.

In order to challenge the vector suite proposed here we selected two "difficult to express" RP like DprE1, and MPK4. For the first protein evaluated (DprE1) the vector suite demonstrated that the expression protein improved when the target protein was fused to Sumo, Trx, DsbC, MBP, and CelDnc solubility enhancer proteins. Among them the best results concerning solubility and quantities of stable protein was achieved when DprE1 was fused to CelDnc and subsequently cleaved by TEV. In the second case, only two out of 12 conditions evaluated were able to express MPK4 in the soluble fraction and only one (pT7-DsbC-MPK4 construct) remains soluble after TEV cleavage. Interestingly, high yield of this fusion construct remained as a decamer before TEV cleavage, so after improving purification protocols (like the use of strepTag II or ion exchange chromatography), the entire fusion can be used for crystallization screenings.

Despite the fact that, many fusion proteins were evaluated, it remains difficult to define a "universal fusion protein." Different options are commercially available (MBP, GST, Trx, DsbC, NusA, etc), and several groups have found new proteins that can be promising alternatives to obtain a soluble and homogeneous recombinant protein (Chatterjee and Esposito, 2006; DelProposto et al., 2009; Cheng et al., 2010; Song et al., 2011) by fusing the target gene. In this work, we evaluated the use of a novel fusion protein, CelDnc that is thermostable (Tm: 71.4°C) and is expressed in massive amounts in *E coli* system. CelD is an endo-β-glucanase (EC 3.2.1.4) from *C. thermocellum* and is part of the cellulose degrading complex termed cellulosome composed of a large number of individual enzymes (Kataeva et al., 1997).

When this protein was evaluated as a solubility fusion enhancer for DprE1 the results showed an increasing solubility performance for this molecule compared with other classical fusion enhancers like MBP. After expression and IMAC purification was done the CelDnc fusion was soluble in large amounts. Moreover, DprE1 was still soluble, monomeric and presented FAD binding properties even after the proteolitical removal of CelDnc demonstrating the utility of this fusion protein that can be taken into account when solubility screening is performed.

In this work we propose a new vector suite and a new fusion enhancer molecule with chances to improve the solubility of different RP. The vector suite proposed here allows the evaluation of five different fusion proteins or only the HisTag in combination with two different promoters, giving rise to 12 different constructs for a single target gene. Altogether, our results suggest that this expression system could be an interesting tool to improve solubility problems of RP.

Moreover, the screening protocol can be further improved. In the present work we used Rosetta cells for the screening of RP production. Different *E. coli* strains can be evaluated in parallel like the use of strains for disulfide bond formation (Shuffle, New Engalnd Biolabs), reduced mRNA degradation (BL21 Star, Invitrogen) among others. Also, the co-expression of chaperones or molecular partners can be included if they are in a vector compatible with a ColE1 replication origin. By the complementation of such variables with the vector suite, a great number of conditions can be screened, increasing the chances of finding the optimal context for target protein production.

It was shown that the sequence at the translation initiation region (TIR) can have a detrimental effect in protein production due to the generation of secondary structures in the messenger RNA that can hamper the translation by the ribosome complex. In this regard a predictive method was developed for designing synthetic ribosome binding sites (RBS) that can minimize the formation of secondary structures at RNA level, so increasing the translation rate (Salis et al., 2009; Salis, 2011). Because the nucleotide sequence from +1 to +25 is the same in all vectors, a new RBS can be designed and introduced into the entire suite increasing translation rates.

Finally, despite the cloning of target genes into the suite was very efficient, false positives were found in some cases. This can be improved, for example, if a toxic gene like the toxin CcdB of type II toxin-antitoxin system is added at the insertion site.

Despite the fact that, more proteins should be tested in this vector suite and that there is no magic formula able to ensure the solubility of different proteins, this could be a useful and economic model to fast track the soluble expression of the RP.

MATERIALS AND METHODS

GENERATION OF THE VECTOR SUITE

For the generation of the vector suite we used a modified version of the pQE80L (Qiagen) as the starter plasmid, that contained a TEV cleavage site after the Histag separated by a GSGS linker (pQE80L-TEV). In a first step we cloned the gene DprE1 into this vector and added the different modules for the vector suite (linkers, strepTag and different fusion proteins) thus generating the T5 series. Then the entire constructs were cloned into the vector pET32a in order to generate the T7 series.

All PCR were done using Phusion polymerase (Finnzymes). For the amplification of the fragments (megaprimer generation) conditions were 30 s at 98°C and 28 cycles of 98°C for 10 s, 59°C for 1 min and 72°C for 1 min with a final extension step at 72°C for 5 min and PCR products were purified by agarose gel. The generated megaprimers contained 30 bp in both ends that overlaps with the insertion site in the destination vectors. The integration into the vectors was done by RF cloning (Unger et al., 2010) and the RF reaction was as follows: 30 s at 98°C and 30 cycles of 98°C for 10 s, 60°C for 1 min and 72°C for 5 min with a final extension step at 72°C for 7 min. For RF reactions 120 ng of megaprimers and 30 ng destination vector were used. 20 μl were digested with 2 μl Fast Digest DpnI (Thermo) for 15 min at 37°C in order to remove parental plasmid, and 5 μl were used to transform 50 μl of competent DH5α *E. coli* cells. Positive clones were confirmed by colony PCR by using Taq polymerase (Invitrogen) with the same primers used for megaprimer generation. Colony PCR was as follows, 95°C for 3 min, 25 cycles of 95°C for 30 s, 60°C for 30 s and 72°C for 2 min followed by a final extension step at 72°C for 5 min. Positive colonies were selected for plasmid extraction and confirmed by sequencing.

The gene for DprE1 was amplified from *M. smegmatis* genomic DNA using the primers QE3790For and QE3790Rev for the generation of the megaprimer (**Table 1**). The product was cloned into the vector pQE80L-TEV by RF cloning to generate the construct pDprE1. The genes coding for CelDwt or the truncated version CelDnc (residues 32–577), were amplified from the plasmid pCT603 (Chaffotte et al., 1992) with the primers CelDwtNFor and CelDwtCRev for CelDwt and primers CelDtruncNFor and CelDtruncCRev for CelDnc (**Table 1**) and cloned by RF in the same vector to generate the constructs pCelD and pCelDnc. The construct pDprE1 was used for the insertion of CelDnc in the 5′ of DprE1 (between the HisTag and the GSGS linker, **Figure 1**). CelDnc was amplified from the pCelDnc construct using primers CelDInsFor and CelDInsRev. The forward primer was designed also to add a GSSG linker to separate the HisTag from the fusion partner generating the construct pCelD-DprE1. The generated constructs (pDprE1 and pCelD-DprE1) were then used to add the last module of the vector, the C-terminal strepTag II. The strepTag II was inserted at the C-terminus separated by a GSGS linker with primers strepCterFor and strepCterRev (**Table 1**) for the generation of the vector pT5-DprE1 (HisTag alone) and pT5-CelD-DprE1 (CelDnc fusion). The primers anneal each other, so they were used without addition of DNA for the generation of the megaprimer. The generated pT5-CelD-DprE1 vector was then used for the insertion and replacement of CelDnc by other fusion partners. In this regard the primers SumoFor and SumoRev; TrxFor, and TrxRev; MBPFor and MBPRev and DsbCFor and DsbCRev were used for the insertion of Sumo, TrxA, MBP, and DsbC, respectively, (**Table 1**). The genes were amplified from *Saccharomyces cerevisiae* for Sumo, pET32a (Novagen) for TrxA, pMAL (New England Biolabs) for MBP; and *E. coli* genome for DsbC. By this way, the T5 vector series was completed. All 6 vectors were confirmed by sequencing with the QEFor and QERev plasmid primers. For the case of MBP and CelDnc constructs internal primers were also used in order to cover the entire sequence. The last step was to transfer the modules into a T7 context. To do this, we selected the pET32a (Novagen) as a destination vectoramplifying the entire cassette from T5 series (from MRGS-HisTag up to the strepTag II for the different fusions) with the primers T5T7For and T5T7Rev and replacing the expression cassette of the pET32a vector. The generated megaprimers were used for the RF reactions. By this way the vector suite was completed containing the gene DprE1 in all 12 vectors for expression screening.

CLONING OF GFP AND MPK4 INTO THE SUITE OF VECTORS

Leishmania major MPK4 gene was amplified with primers MPK4For and MPK4Rev from a pGem vector containing the gene. GFP was amplified with primers GFPFor and GFPRev from a pET vector containing a GFP variant that is well expressed in *E. coli* (Waldo et al., 1999).

The 12 vectors were added to 12 different PCR tubes, and the amplified products were used as megaprimers for the RF reaction using the HF buffer from Phusion polymerase. After digestion of 20 μl PCR products with 2 μl DpnI, chemical competent cells were transformed with 5 μl RF reaction in a PCR machine with the following program: 30 min at 4°C, 45 s at 42°C, 3 min at 4°C, addition of 100 μl of LB, 1 h at 37°C, and plating of 100 μl in agar plates containing ampicillin. Four colonies for each construct were selected and confirmed by colony PCR and sequenced. After the analysis we found out that in most cases all were positive (or at least three of four were positive) giving a percentage of success of more than 80%.

EXPRESSION SCREENING OF GFP AND DprE1

Chemocompetent Rosetta-pLysS cells were transformed with 5 μl of purified plasmids as described above and then incubated in a shaker ON at 37°C in 1 ml of LB with chloramphenicol and ampicillin in a 96 × deep-well plate. 100 μl of ON culture were used to inoculate 4 ml of Terrific Broth in 24 × deep-well plates by duplicate. Cultures were incubated at 37°C until $D.O._{600}$ reached 1.0–1.2. At that moment one plate was induced with 1 mM IPTG and left at 37°C for 4 h. The other 24 deep-well was incubated at 17°C for 15 min to cooling it and then induced with 1 mM IPTG ON at the same temperature. After induction time was reached, cells were harvested, resuspended in 1 ml lysis buffer (50 mM Tris pH 8.0; 300 mM NaCl, 10 mM imidazol, 0.5 mg/ml lysozyme) and frozen at −80°C. After thawing cells, 10 units of DNase I and 10 μl of 2M $MgSO_4$ were added and incubated with shaking for 20 min at 20°C. Then 200 μl of Nickel beads (Qiagen) equilibrated in binding buffer (50 mM Tris pH 8.0; 300 mM NaCl, 10 mM imidazol) were added to cell extracts and incubated for 15 min at 20°C. Cell extracts were then transferred to a 96×-well filter plate assembled in a vacuum device, and bound protein was washed with 2 ml of binding buffer. An additional wash step was done with 2 ml of binding buffer containing 50 mM imidazol. Elution was done with 160 μl of elution buffer [50 mM Tris pH 8.0; 300 mM NaCl, 500 mM imidazol; for a detailed protocol, see (Saez and Vincentelli, 2013)]. Eluates were divided in two groups for evaluation of uncleaved protein and assessment of TEV cleavage ON at 18°C.

Samples were then loaded into an E-PAGE 96 acrylamide gel (Invitrogen).

EXPRESSION SCREENING OF MPK4

Expression screening and purification of MPK4 constructs was made in a similar way than for GFP and DprE1 but only 17°C of induction was evaluated. Purification steps were the same but the pipeting scheme was done automatically by using a TECAN Freedom EVO®200. Expression analysis was done also automatically by using a Labchip GX II (Caliper, USA) microfluidic detection system.

LARGE SCALE EXPRESSION AND PURIFICATION OF DsbC-MPK4

DsbC-MPK4 was expressed in Terrific Broth (TB) supplemented with ampicillin and chloramphenicol and induction was done at $D.O_{600}$: 1.2 ON at 17°C with 1 mM IPTG. Pellets were resuspended in lysis buffer and frozen at −80°C. After thawing, the pellets were sonicated and centrifugated at 15.000 × *g*. Soluble fraction was injected in a 1 ml IMAC column (GE Healthcare) equilibrated in binding buffer. Elution was done in a linear gradient of 5–100% B in 10 column volumes (CV) with elution buffer. Purified protein was cleaved with TEV protease in a 1:30 protein:enzyme ratio and dialyzed against cleavage buffer (50 mM Tris pH 8.0; 150 mM NaCl, 1 mM DTT) ON at 8°C. Sample was filtered through 0.22 μm to remove precipitates, and analyzed by SDS-PAGE.

EXPRESSION AND PURIFICATION OF CelD AND CelDnc

Production of CelD and CelDnc was done in M15pREP4 from the constructs pCelD and pCelDnc, respectively, in 1 l 2YT supplemented with ampicillin and kanamycin, and induced with 1 mM IPTG at D.O. 1.0 ON at 37°C. IMAC was done like for the case of DsbC-MPK4 but using a 5 ml column and only half of the soluble fraction was used. TEV cleavage was done as before and desalted in order to remove imidazole. The reaction was injected in a second IMAC under same conditions as above and the flow through containing the cleaved protein was injected in a Superdex 200 16/60 (GE Healthcare) equilibrated with buffer 40 mM Tris pH 7.7.

DSC ANALYSIS OF CelD AND CelDnc

Differential scanning calorimetry (DSC) experiments were carried out in PBS, in a VP-DSC instrument (Microcal, Northampton, MA, USA) and data analyzed with the software supplied with the equipment. The temperature was increased at 1°C per minute from 30 to 80°C, and proteins were added at concentration of 1 mg/ml for CelD and CelDnc.

LARGE SCALE EXPRESSION AND PURIFICATION OF pT5-DprE1, pT5-CelD-DprE1 AND pT5-MBP-DprE1

Induction of p5DprE1, p5CelDnc-DprE1 and p5MBP-DprE1 were done in M15pREP4 with 1 mM IPTG in 1 l 2YT supplemented with ampicillin (100 μg/ml), kanamycin (50 μg/ml) and 15 μM FAD at D.O.: 1.0–1.2 during 4 h at 37°C. Cells were harvested, resuspended in lysis buffer and frozen at −80°C. After thawing the cells, were lysed and protein purified as before. Purified protein was cleaved with TEV in a 1:30 ratio, and dialysed against cleavage buffer. The product was then purified by a second IMAC and injected in a Superdex 200 16/60 equilibrated with buffer 25 mM Tris pH 8.0; 150 mM NaCl.

ACKNOWLEDGMENTS

This work was partially funded by FOCEM (MERCOSUR Structural Convergence Fund), COF 03/11 and CYTED Program. Agustín Correa was supported by a doctoral program of the Agencia Nacional de Investigación e Innovación, Uruguay. We wish to thank Dr. Trajtemberg and Sofía Horjales from the Crystallography Unit (PXF) of the Institut Pasteur de Montevideo for giving the plasmid pQE80L-TEV and pGem-MPK4 and Mrs. Natalia López for helpful secretarial assistance.

REFERENCES

Aslanidis, C., and de Jong, P. J. (1990). Ligation-independent cloning of PCR products (LIC-PCR). *Nucleic Acids Res.* 18, 6069–6074. doi: 10.1093/nar/18.20.6069

Berrow, N. S., Alderton, D., Sainsbury, S., Nettleship, J., Assenberg, R., Rahman, N., et al. (2007). A versatile ligation-independent cloning method suitable for high-throughput expression screening applications. *Nucleic Acids Res.* 35, e45. doi: 10.1093/nar/gkm047

Cabrita, L. D., Dai, W., and Bottomley, S. P. (2006). A family of *E. coli* expression vectors for laboratory scale and high throughput soluble protein production. *BMC Biotechnol.* 6:12. doi: 10.1186/1472-6750-6–12

Chaffotte, A. F., Guillou, Y., and Goldberg, M. E. (1992). Inclusion bodies of the thermophilic endoglucanase D from *Clostridium thermocellum* are made of native enzyme that resists 8 M urea. *Eur. J. Biochem.* 205, 369–373. doi: 10.1111/j.1432-1033.1992.tb16789.x

Chatterjee, D. K., and Esposito, D. (2006). Enhanced soluble protein expression using two new fusion tags. *Protein Expr. Purif.* 46, 122–129. doi: 10.1016/j.pep.2005.07.028

Cheng, Y., Gu, J., Wang, H. G., Yu, S., Liu, Y. Q., Ning, Y. L., et al. (2010). EspA is a novel fusion partner for expression of foreign proteins in *Escherichia coli*. *J. Biotechnol.* 150, 380–388. doi: 10.1016/j.jbiotec.2010.09.940

Correa, A., and Oppezzo, P. (2011). Tuning different expression parameters to achieve soluble recombinant proteins in *E. coli*: advantages of high-throughput screening. *Biotechnol. J.* 6, 715–730. doi: 10.1002/biot.201100025

Curiel, J. A., De Las Rivas, B., Mancheno, J. M., and Munoz, R. (2010). The pURI family of expression vectors: a versatile set of ligation independent cloning plasmids for producing recombinant His-fusion proteins. *Protein Expr. Purif.* 76, 44–53. doi: 10.1016/j.pep.2010.10.013

Dahlroth, S. L., Nordlund, P., and Cornvik, T. (2006). Colony filtration blotting for screening soluble expression in *Escherichia coli*. *Nat. Protoc.* 1, 253–258. doi: 10.1038/nprot.2006.39

DelProposto, J., Majmudar, C. Y., Smith, J. L., and Brown, W. C. (2009). Mocr: a novel fusion tag for enhancing solubility that is compatible with structural biology applications. *Protein Expr. Purif.* 63, 40–49. doi: 10.1016/j.pep.2008.08.011

Esposito, D., Garvey, L. A., and Chakiath, C. S. (2009). Gateway cloning for protein expression. *Methods Mol. Biol.* 498, 31–54. doi: 10.1007/978-1-59745-196-3_3

Kapust, R. B., Tozser, J., Copeland, T. D., and Waugh, D. S. (2002). The P1' specificity of tobacco etch virus protease. *Biochem. Biophys. Res. Commun.* 294, 949–955. doi: 10.1016/S0006-291X(02)00574-0

Kapust, R. B., and Waugh, D. S. (1999). *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci.* 8, 1668–1674. doi: 10.1110/ps.8.8.1668

Kataeva, I., Guglielmi, G., and Beguin, P. (1997). Interaction between *Clostridium thermocellum* endoglucanase CelD and polypeptides derived from the cellulosome-integrating protein CipA: stoichiometry and cellulolytic activity of the complexes. *Biochem. J.* 326 (Pt 2), 617–624.

LaVallie, E. R., Lu, Z., Diblasio-Smith, E. A., Collins-Racie, L. A., and McCoy, J. M. (2000). Thioredoxin as a fusion partner for production of soluble recombinant proteins in *Escherichia coli*. *Methods Enzymol.* 326, 322–340. doi: 10.1016/S0076-6879(00)26063-1

Luna-Vargas, M. P., Christodoulou, E., Alfieri, A., Van Dijk, W. J., Stadnik, M., Hibbert, R. G., et al. (2011). Enabling high-throughput ligation-independent cloning and protein expression for the family of ubiquitin specific proteases. *J. Struct. Biol.* 175, 113–119. doi: 10.1016/j.jsb.2011.03.017

Magnusdottir, A., Johansson, I., Dahlgren, L. G., Nordlund, P., and Berglund, H. (2009). Enabling IMAC purification of low abundance recombinant proteins from *E. coli* lysates. *Nat. Methods* 6, 477–478. doi: 10.1038/nmeth0709-477

Marblestone, J. G., Edavettal, S. C., Lim, Y., Lim, P., Zuo, X., and Butt, T. R. (2006). Comparison of SUMO fusion technology with traditional gene fusion systems: enhanced expression and solubility with SUMO. *Protein Sci.* 15, 182–189. doi: 10.1110/ps.051812706

Murphy, M. B., and Doyle, S. A. (2005). High-throughput purification of hexahistidine-tagged proteins expressed in *E. coli*. *Methods Mol. Biol.* 310, 123–130. doi: 10.1007/978-1-59259-948-6_9

Neres, J., Pojer, F., Molteni, E., Chiarelli, L. R., Dhar, N., Boy-Rottger, S., et al. (2012). Structural basis for benzothiazinone-mediated killing of *Mycobacterium tuberculosis*. *Sci. Transl. Med.* 4, 150ra121. doi: 10.1126/scitranslmed.3004395

Nozach, H., Fruchart-Gaillard, C., Fenaille, F., Beau, F., Ramos, O. H., Douzi, B., et al. (2013). High throughput screening identifies disulfide isomerase DsbC as a very efficient partner for recombinant expression of small disulfide-rich proteins in *E. coli*. *Microb. Cell Fact.* 12, 37. doi: 10.1186/1475-2859-12–37

Saez, N. J., and Vincentelli, R. (2013). High-throughput expression screening and purification of recombinant proteins in *E. coli*. *Methods Mol. Biol.* 1091, 33–53. doi: 10.1007/978-1-62703-691-7_3

Salis, H. M. (2011). The ribosome binding site calculator. *Methods Enzymol.* 498, 19–42. doi: 10.1016/B978-0-12-385120-8.00002–4

Salis, H. M., Mirsky, E. A., and Voigt, C. A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.* 27, 946–950. doi: 10.1038/nbt.1568

Schafer, F., Romer, U., Emmerlich, M., Blumer, J., Lubenow, H., and Steinert, K. (2002). Automated high-throughput purification of 6xHis-tagged proteins. *J. Biomol. Tech.* 13, 131–142.

Schmidt, T. G., and Skerra, A. (2007). The Strep-tag system for one-step purification and high-affinity detection or capturing of proteins. *Nat. Protoc.* 2, 1528–1535. doi: 10.1038/nprot.2007.209

Song, J. A., Lee, D. S., Park, J. S., Han, K. Y., and Lee, J. (2011). A novel *Escherichia coli* solubility enhancer protein for fusion expression of aggregation-prone heterologous proteins. *Enzyme Microb. Technol.* 49, 124–130. doi: 10.1016/j.enzmictec.2011.04.013

Steen, J., Uhlen, M., Hober, S., and Ottosson, J. (2006). High-throughput protein purification using an automated set-up for high-yield affinity chromatography. *Protein Expr. Purif.* 46, 173–178. doi: 10.1016/j.pep.2005.12.010

Unger, T., Jacobovitch, Y., Dantes, A., Bernheim, R., and Peleg, Y. (2010). Applications of the restriction free (RF) cloning procedure for molecular manipulations and protein expression. *J. Struct. Biol.* 172, 34–44. doi: 10.1016/j.jsb.2010.06.016

van den Berg, S., Lofdahl, P. A., Hard, T., and Berglund, H. (2006). Improved solubility of TEV protease by directed evolution. *J. Biotechnol.* 121, 291–298. doi: 10.1016/j.jbiotec.2005.08.006

Vincentelli, R., Cimino, A., Geerlof, A., Kubo, A., Satou, Y., and Cambillau, C. (2011). High-throughput protein expression screening and purification in *Escherichia coli*. *Methods* 55, 65–72. doi: 10.1016/j.ymeth.2011.08.010

Vincentelli, R., and Romier, C. (2013). Expression in *Escherichia coli*: becoming faster and more complex. *Curr. Opin. Struct. Biol.* 23, 326–334. doi: 10.1016/j.sbi.2013.01.006

Waldo, G. S., Standish, B. M., Berendzen, J., and Terwilliger, T. C. (1999). Rapid protein-folding assay using green fluorescent protein. *Nat. Biotechnol.* 17, 691–695. doi: 10.1038/10904

Young, C. L., Britton, Z. T., and Robinson, A. S. (2012). Recombinant protein expression and purification: a comprehensive review of affinity tags and microbial applications. *Biotechnol. J.* 7, 620–634. doi: 10.1002/biot.201100155

Yumerefendi, H., Tarendeau, F., Mas, P. J., and Hart, D. J. (2010). ESPRIT: an automated, library-based method for mapping and soluble expression of protein domains from challenging targets. *J. Struct. Biol.* 172, 66–74. doi: 10.1016/j.jsb.2010.02.021

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 21 December 2013; paper pending published: 13 January 2014; accepted: 05 February 2014; published online: 25 February 2014.

*Citation: Correa A, Ortega C, Obal G, Alzari P, Vincentelli R and Oppezzo P (2014) Generation of a vector suite for protein solubility screening. Front. Microbiol. **5**:67. doi: 10.3389/fmicb.2014.00067*

This article was submitted to Microbiotechnology, Ecotoxicology and Bioremediation, a section of the journal Frontiers in Microbiology.

2. Aplicación y caracterización de una nueva librería de Afitinas como inhibidores enzimáticos de glicosidasas.

"Potent and specific inhibition of glycosidases by small artificial binding proteins (Affitins)"

Agustín Correa[a,b,1], Sabino Pacheco[b,c,1], Ariel Mechaly[b,2], Gonzalo Obal[d], Ghislaine Béhar[c], Barbara Mouratou[c], Pablo Oppezzo[a], Pedro M. Alzari[b] and Frédéric Pecorari[c,*]

[a]Institut Pasteur de Montevideo, Recombinant Protein Unit, Mataojo 2020, Montevideo CP 11400, Uruguay.

[b]Institut Pasteur, Unité de Microbiologie Structurale, CNRS UMR 3528, 25 rue du Dr. Roux, 72724 Paris Cedex 15, France.

[c]University of Nantes, CRCNA - UMR 892 INSERM , 6299 CNRS, 8 quai Moncousu, BP 70721, 44007 Nantes, Cedex 1, France.

[d]Institut Pasteur de Montevideo, Protein Biophysics Unit, Mataojo 2020, Montevideo CP 11400, Uruguay.

[1]These authors contributed equally to this work

[2]Present address: Institut Pasteur de Montevideo, Unit of Protein Crystallography, Mataojo 2020, Montevideo CP 11400, Uruguay

[*]Address for correspondence: Université de Nantes, CRCNA - UMR 892 INSERM , 6299 CNRS, 8 quai Moncousu, BP 70721, 44007 Nantes Cedex 1, France. Tel: 33-2 40 41 28 51; E-mail: frederic.pecorari@univ-nantes.fr

La generación de inhibidores enzimáticos es de gran importancia no sólo para la investigación básica sino también para el desarrollo de aplicaciones biomédicas. En este sentido, el uso de pequeñas moléculas ha tenido éxito en el tratamiento contra diversos tipos de cáncer como en el caso de inhibidores de varias tirosina quinasas o de la polimerasa poly(ADP-ribosa)polimerasa (PARP) (7). Sin embargo, estas moléculas pueden también inhibir otras enzimas relacionadas funcionalmente a la proteína blanco, que poseen sitios activos estructuralmente similares y que están presentes en el mismo organismo llevando a una promiscuidad de clase y por ende limitando sus aplicaciones. Este es por ejemplo el caso de las glicósido hidrolasas o glicosidasas, proteínas que han evolucionado de un ancestro común dando lugar a especificidades funcionales diferentes pero que presentan mecanismos catalíticos y sitios activos similares, haciendo difícil su inhibición específica (267). Las glicosidasas están implicadas en varios desórdenes metabólicos y enfermedades humanas como ser la diabetes de tipo 2, la enfermedad de Gaucher, el cáncer y el asma (227, 268-270). Debido a las similitudes estructurales entre las diferentes glicosidasas, pocos inhibidores útiles para terapias basados en pequeñas moléculas han sido desarrollados contra estas enzimas (271).

En este sentido el uso de inhibidores proteicos como alternativa a las pequeñas moléculas puede ser una estrategia interesante, ya que al ofrecer superficies de unión más extensas, pueden presentar mayor especificidad al reconocer también residuos en las cercanías de los sitios activos que no estén conservados. Moléculas como ser los AcMo pueden ser una opción válida con dichos fines, sin embargo se ha hallado que no es común que reconozcan superficies cóncavas (10) como las típicamente encontradas en los sitios activos de las glicosidasas (267). Es así que el uso de proteínas de unión diferentes a los anticuerpos puede ser una alternativa muy interesante para inhibir este tipo de enzimas. Además, una estrategia conveniente y general puede ser la de desarrollar inhibidores, que a partir de un mismo plagado proteico o scaffold, puedan reconocer diferentes topologías estructurales al unir y cubrir superficies planas, o reconocer y penetrar en sitios activos profundos.

De esta manera es que explotamos la plasticidad y estabilidad de las proteínas de unión conocidas como afitinas, derivadas de la proteína de 7 kDa Sac7d de la arquea *Sulfolobus acidocaldarius*. Para lograr obtener proteínas de unión que puedan reconocer y unir topologías diferentes en una proteína, es que se utilizó además de la librería ya descripta (L1 y L2) para las afitinas (11, 192), una librería con un nuevo diseño (L3 y L4) donde el loop 2 que conecta las hebras β 3 y 4 de Sac7d se extendió con 4 residuos aleatorios. Mientras en el caso de L1, la superficie de unión esta compuesta por 14 sitios aleatorios formando una superficie plana, L2

posee 10 sitios aleatorios formando una superficie plana con un loop (192). Como prueba de concepto de que con ambas librerías, diferentes topologías pueden ser reconocidas y que son capaces además, de reconocer de forma específica glicosidasas, es que se seleccionaron afitinas mediante ribosome display (RD) contra la endo-glicosidasa CelD de *C. thermocellum* (EC 3.2.1.4). También se analizó la afitina previamente obtenida contra la endo-glicosidasa lisozima, de huevo de gallina (EC 3.2.1.17) (12). Estas glicosidasas además tienen la ventaja que se expresan fácilmente o están accesibles comercialmente, son propensas a cristalizar y para el caso de CelD es activa incluso a 60˚C, permitiéndonos evaluar la unión/inhibición a distintas temperaturas y aumentando las probabilidades de éxito para estudios estructurales.

Luego de realizar 4 rondas de RD contra CelD, se observó enriquecimiento únicamente para la librería conteniendo el loop randomizado. Si bien con la librería sin el loop se continuó hasta la sexta ronda, no se observó enriquecimiento de proteínas de unión. De esta forma fueron aislados clones provenientes de la nueva librería obteniéndose 16 secuencias únicas, donde se vio un motivo conservado dentro del loop, Leu-Thr/Ser-Lys excepto por la afitina H3 donde la Leu esta cambiada a Asn y la Lys a Arg. Se seleccionaron las afitinas H3 y E12 para hacer una producción a mayor escala (1 lt). Al analizar la actividad de CelD en presencia de cada una de las 2 afitinas, se observó inhibición de la actividad de CelD, incluso a 60˚C. Utilizando la afitina obtenida contra la lisozima (H4), también se evaluó si además de unir a esta, era capaz de inhibir a la enzima, encontrándose inhibición de la lisozima. Se determinó la Ki para H4 y dos de las afitinas anti-CelD (E12 y H3), determinándose que eran inhibidores potentes con una Ki de 45, 95 y 111 nM respectivamente. Las estabilidades térmicas de las 3 afitinas se determinaron mediante DSC (Calorimetría de barrido diferencial) encontrándose que las tres afitinas son termostables con una temperatura de fusión o desnaturalización entre 67-81°C. Las afinidades de unión fueron determinadas mediante ITC (Calorimetría isotérmica de titulación), mostrando un modo de unión simple 1:1, con una afinidad de 11, 48 y 98 nM para H4, H3 y E12 respectivamente. Más aún para las afitinas anti-CelD también se determino la afinidad por ITC a 60˚C, donde no se observaron cambios importantes en las mismas. Esto muestra que si bien las selecciones fueron realizadas a 4˚C, es posible obtener afitinas capaces de unir/inhibir sus blancos en un amplio rango de temperaturas. Mediante ELISA e ITC, se evaluó la reacción cruzada entre las afitinas y las dos glicosidasas en estudio encontrándose que estas unían específicamente al blanco para el que fueron seleccionadas.

Para realizar estudios estructurales con los diferentes complejos afitina-glicosidasa se realizó un cribado de 384 condiciones de cristalización para los complejos H4-lisozima y H3-CelD. Se obtuvieron cristales para los 2 complejos y se determinaron las correspondientes estructuras

tridimensionales a alta resolución (1.5 y 1.6 Å para H4-lisozima y H3-CelD respectivamente). La condición de cristalización para el complejo H3-CelD se utilizó como punto de partida para cristalizar el complejo E12-CelD en condiciones similares. De esta forma se logró obtener la estructura de E12-CelD a 1.1 Å de resolución. Estos datos representan las primeras estructuras obtenidas para complejos entre afitinas y proteínas blanco. El análisis estructural demostró que, para el caso de las afitinas anti-CelD, el loop randomizado es el principal responsable de la interacción dado que penetra en la cavidad del sitio activo de CelD e interacciona con residuos catalíticamente importantes, impidiendo el acceso del sustrato. En el caso H4-lisozima, el modo de unión es diferente, con una interface relativamente plana que bloquea el sitio activo de la glicosidasa.

De esta manera, se logró obtener afitinas termostables que unen con alta afinidad y especificidad las dos glicosidasas, inhibiéndolas de forma potente mediante modos de interacción diferentes. Más aún, para el caso de CelD se determinó un efecto inhibitorio incluso a una temperatura 15 veces mayor a la que fueron seleccionadas las afitinas confirmando su versatilidad. Finalmente, el nuevo diseño de librerías mostró ser muy útil para la unión de enzimas con cavidades profundas ya que sólo se obtuvieron afitinas que inhibían a CelD, sin haber hecho una selección específica para ello. Probablemente esto no sea un caso fortuito, lo que puede indicar que este tipo de diseño tenga cierta tendencia a interaccionar con superficies cóncavas.

Esta segunda etapa de la Tesis culminó con el artículo recientemente aceptado en la revista **Plos One** en donde se propone que las afitinas pueden ser seleccionadas como inhibidores enzimáticos específicos no sólo contra glicosidasas sino posiblemente contra otros tipos de enzimas en general.

OPEN ACCESS Freely available online

PLOS ONE

Potent and Specific Inhibition of Glycosidases by Small Artificial Binding Proteins (Affitins)

Agustín Correa[1,2◑], Sabino Pacheco[2,3,4,5◑], Ariel E. Mechaly[2¤], Gonzalo Obal[6], Ghislaine Béhar[3,4,5], Barbara Mouratou[3,4,5], Pablo Oppezzo[1], Pedro M. Alzari[2], Frédéric Pecorari[3,4,5]*

1 Institut Pasteur de Montevideo, Recombinant Protein Unit, Montevideo, Uruguay, **2** Institut Pasteur, Unité de Microbiologie Structurale, CNRS UMR 3528, Paris, France, **3** INSERM UMR 892 - CRCNA, Nantes, France, **4** CNRS UMR 6299, Nantes, France, **5** University of Nantes, Nantes, France, **6** Institut Pasteur de Montevideo, Protein Biophysics Unit, Montevideo, Uruguay

Abstract

Glycosidases are associated with various human diseases. The development of efficient and specific inhibitors may provide powerful tools to modulate their activity. However, achieving high selectivity is a major challenge given that glycosidases with different functions can have similar enzymatic mechanisms and active-site architectures. As an alternative approach to small-chemical compounds, proteinaceous inhibitors might provide a better specificity by involving a larger surface area of interaction. We report here the design and characterization of proteinaceous inhibitors that specifically target endoglycosidases representative of the two major mechanistic classes; retaining and inverting glycosidases. These inhibitors consist of artificial affinity proteins, Affitins, selected against the thermophilic CelD from *Clostridium thermocellum* and lysozyme from hen egg. They were obtained from libraries of Sac7d variants, which involve either the randomization of a surface or the randomization of a surface and an artificially-extended loop. Glycosidase binders exhibited affinities in the nanomolar range with no cross-recognition, with efficient inhibition of lysozyme ($K_i = 45$ nM) and CelD ($K_i = 95$ and 111 nM), high expression yields in *Escherichia coli*, solubility, and thermal stabilities up to 81.1°C. The crystal structures of glycosidase-Affitin complexes validate our library designs. We observed that Affitins prevented substrate access by two modes of binding; covering or penetrating the catalytic site *via* the extended loop. In addition, Affitins formed salt-bridges with residues essential for enzymatic activity. These results lead us to propose the use of Affitins as versatile selective glycosidase inhibitors and, potentially, as enzymatic inhibitors in general.

Citation: Correa A, Pacheco S, Mechaly AE, Obal G, Béhar G, et al. (2014) Potent and Specific Inhibition of Glycosidases by Small Artificial Binding Proteins (Affitins). PLoS ONE 9(5): e97438. doi:10.1371/journal.pone.0097438

Editor: Mark J. van Raaij, Centro Nacional de Biotecnologia - CSIC, Spain

Received February 23, 2014; **Accepted** April 17, 2014; **Published** May 13, 2014

Funding: S.P. was supported by a post-doctoral fellowship of the Institut Pasteur (Paris, France). A.C. was supported by a doctoral program of the Agencia Nacional de Investigación e Innovación, Uruguay. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have read the journal's policy and have the following conflicts. P.M.A. and F.P. are inventors of a patent application, owned by Institut Pasteur and Centre National de la Recherche Scientifique (CNRS): ''OB-fold used as scaffold for engineering new specific binders''; PCT/IB2007/004388, that includes some of the ideas described in this manuscript. F.P. is a cofounder of Affilogic, a spin-off company of Institut Pasteur and CNRS, which has a license agreement related to this application. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: frederic.pecorari@univ-nantes.fr

◑ These authors contributed equally to this work.

¤ Current address: Institut Pasteur de Montevideo, Unit of Protein Crystallography, Montevideo, Uruguay

Introduction

Glycosidases are involved in a variety of metabolic disorders and human diseases such as type II diabetes, Gaucher disease, cancers and asthma [1,2,3,4]. They are thus actively studied not only to probe their functions, but also as targets for inhibitor drugs to treat human diseases. However, achieving specific and efficient inhibition of a particular glycosidase represents a major challenge because a given organism can produce many different glycosidases, and also because this class of enzymes has evolved different functional specificities from a single structural scaffold, giving rise to similar active-site architectures and catalytic mechanisms. *In vivo*, a lack of selectivity for a drug can increase the risk of undesirable effects or even lead to toxicity [5] by off-target effects.

The use of small-molecular weight compounds is a powerful approach to modulate the activity of individual glycosidases [6], and a number of small-molecule inhibitors have been described for these enzymes. Although this class of inhibitors is attractive for the development of drugs, they can interact with non-target proteins and thus few high-quality inhibitors useful for therapy have been reported (for a review see refs. [6] and [7]). An alternative strategy is the development of proteinaceous inhibitors. Compared to small-molecule ligand-protein interactions, protein-protein or protein-antibody interactions generally involve much larger interfaces (typically 800–1000 $Å^2$, [8,9,10]), a favorable feature to achieve binding with high specificity and selectivity. Antibodies can bind quite different compounds specifically, but it may be difficult to obtain candidates that bind a cleft-shaped active site [11], such as those of endo-glycosidases. Alternatives to classic antibodies have emerged based on immunoglobulin or non-immunoglobulin folds (for a review see refs. [12] and [13]) to derive specific binders of targeted proteins. Only a few of these binders have been shown to be potent enzymatic inhibitors and even fewer have been described at the structural level to

understand their mechanisms of inhibition. As examples, the Ecotin scaffold has been used to generate a highly specific inhibitor of the protease kallikrein with a K_i of 11 pM [14] while binders with inhibition properties for hen egg white lysozyme (HEWL) have been derived from various proteins (VHH, shark IgNAR and an anticodon recognition domain of the aspartyl tRNA synthetase), and have been structurally described to mimic the oligosaccharide substrate of this glycosidase [11,15,16,17,18].

A general and convenient strategy to develop inhibitors would be to use a unique scaffold protein able to either cover or deeply penetrate active sites. The success of this approach depends essentially on the ability of the scaffold protein to recognize catalytic sites with different shapes. As an important step towards this goal, we have exploited the plasticity and stability of artificial 7 kDa affinity proteins (Affitins) [19,20,21,22,23] derived from extremophilic proteins, such as DNA-binding protein 7d (Sac7d), which are found in various Archaea such as *Sulfolobus*, *Acidianus*, and *Metallospharea* genera. With their small size and their low structural complexity, Affitins occupy an intermediate position between peptides and proteins. Previously, we reported that Affitins can bind different epitopes of the same target *via* two different modes of binding: one involving a flat surface and the other involving a flat surface and two short loops [23].

Based on these results, in this work we designed two Affitin libraries in which a loop of Sac7d was extended by four additional randomized residues. As a proof of concept that Affitins may inhibit different glycosidases specifically, we used these libraries (L3 and L4) and those we had previously designed without an extended loop (L1 and L2) to select Affitins specific for the inverting endo-glycosidase CelD from *Clostridium thermocellum* (EC 3.2.1.4). We also analyzed an Affitin specific for the well-studied (retaining endo-glycosidase) HEWL (EC 3.2.1.17) previously selected from the library L1 [20,24]. These two glycosidases hydrolyze the O-glycosyl bond and are representative of the two main glycosidase mechanisms of action [25]. Isolated Affitins were shown to be potent inhibitors of CelD and of HEWL, with K_i in the nanomolar range, without cross-recognition. The crystal structures of Affitin-CelD and Affitin-HEWL complexes revealed their inhibition mechanisms, and provided useful hints for further inhibitor improvement. These results lead us to propose the use of Affitins as versatile and thermostable selective glycosidase inhibitors.

Materials and Methods

Chemicals were purchased from Sigma-Aldrich. Enzymes and buffers for molecular biology were purchased from Thermo Scientific or New England Biolabs unless otherwise indicated. Oligonucleotides were purchased from Eurofins. All PCR were performed using Vent polymerase.

Construction of Libraries and Selections

Since we have observed that two tryptophans at positions 8 and 9 can promote multimerization of Affitins, we either did not randomize these two positions (library L3) or limited their randomization using NHK codons (library L4) that do not encode tryptophan. This codon sub-set also excludes Gly, Cys and Arg. The other positions were randomized using NNS triplets that encode all amino acids and only one stop-codon.

The generation of libraries L1 and L2, which corresponds to the random mutagenesis of positions 7, 8, 9, 21, 22, 24, 26, 29, 31, 33, 40, 42, 44, and 46 and of positions 26, 27, 28, 29, 31, 42, 44, 46, 47, and 48, respectively, in Sac7d protein has been previously described [19,23]. To construct library L3, which corresponds to the random mutagenesis of positions 7, 26, 27, 27a, 27b, 27c, 27d, 28, 29, 31, 44, 46, and 48 in Sac7d protein, the same protocol was used with the following oligonucleotides: T7B (5′- ATACGAAATTAATACGACTCACTATAGGGAGACCACAACGG-3′), T7C (5′-ATACGAAATTAATACGACTCACTATAGGGAGACCACAACGGTTTCCCTC-3′), SDA_MRGS (5′- AGACCACAACGGTTTCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATATCCATGAGAGGATCG-3′), SClib2.1 (5′-GGAGATATATCCATGAGAGGATCGCATCACCATCACCATCACGGATCCGTCAAGGTGAAATTC-3′), SClib6.2.

(5′-GGATCCGTCAAGGTGAAATTCNNSTATAAAGGCGAAGAAAAAGAAGTGGACACTAGTAAGATC-3′), SClib8.3 (5′-CTTGCCGTTGTCGTCGTAGGTAAASNNCACSNNSNNSNNSNNSNNSNNSNNACGCCAAACTTTCTTGATCTTACTAGTGTCCACTTC-3′), SClib6.4.

5′-TAATAACTCTTTCGGGGCATCSNNCTCSNNCACSNNGCCACGGCCGGTCTTGCCGTTGTCGTCGTAGG-3′), SClib2.5 (5′-CCATATAAAGCTTTTTCTCGCGTTCCGCACGCGCTAACATATCTAATAACTCTTTCGGGGCATC-3′), tolAk (5′-CCGCACACCAGTAAGGTGTGCGGTTTCAGTTGCCGCTTTCTTTCT-3′). To construct library L4 which corresponds to the random mutagenesis of positions 7, 8, 9, 26, 27, 27a, 27b, 27c, 27d, 28, 29, 31, 44, 46, and 48 in Sac7d protein, the same protocol was used but replacing SClib6.2 with SClib7.2 (5′-GGATCCGTCAAGGTGAAATTCNNSNHKNHKGGCGAAGAAAAAGAAGTGGACACTAGTAAGATC-3′). Both libraries were constructed in the ribosome display format with estimated numbers of independent variants of about 10^{12} [26].

The preparation of targets by *in vitro* biotinylation was performed as previously described [19,23]. The ribosome display selections were also performed as previously described [26], except that the incubation time for the translation reaction was 10 min while the incubation times for the pre-panning and panning steps were 30 min in both cases. The RT-PCR was as follows: for selection rounds 1 and 2, an initial denaturation step at 95°C for 30 s, followed by 45 cycles of 30 s at 95°C, 30 s at 63°C, and 30 s at 72°C with a final elongation step of 5 min at 72°C. For selection rounds 3 and 4, it was the same program but with 40 cycles instead of 45. For the selections, 100 µl of biotinylated CelD (250 nM for round 1, 200 nM for round 2 and 150 nM for rounds 3 and 4) was bound on MaxiSorp ELISA plates (Nunc) previously coated with NeutrAvidin (Thermo Scientific) or streptavidin (Sigma-Aldrich), which were alternated during four or six selection rounds. To isolate high-affinity binders, the time in wash-steps was increased during the selection (6 washes of 30 s, 1 min, 3 min and 10 min for rounds 1, 2, 3 and 4, respectively).

Analysis of Selected Pools and Isolated Clones

To assess enrichments of the selections, the output RNA obtained after four or five rounds of selection were translated *in vitro* and tested in MaxiSorp ELISA wells coated with streptavidin/NeutrAvidin and biotinylated CelD as previously described [23,26]. A negative control was performed with wells coated with only streptavidin or NeutrAvidin. For anti-CelD Affitins, the RT-PCR product from L3 and L4 obtained after round 4, and which gave a positive signal in ELISA, was cloned into the *Bam*HI and *Hind*III restriction sites of the pFP1001 vector, and the ligation mixture was transformed into *E. coli* DH5αF'IQ (Invitrogen) for the isolation of individual clones [26]. The screening of individual clones was performed by ELISA as described before [23,26].

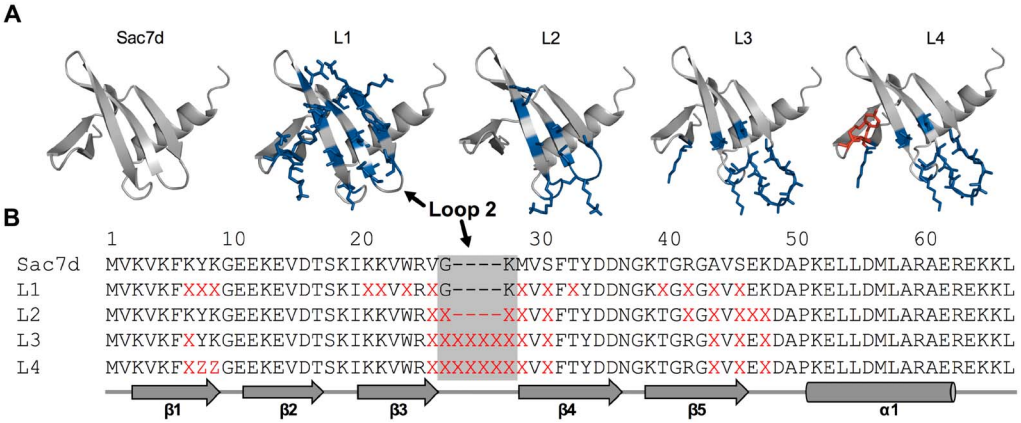


Figure 1. Schematic representation of Affitin libraries. (**A**) Sac7d wild-type structure. Two β-sheets composed of two ($\beta_1\beta_2$) and three ($\beta_3\beta_4\beta_5$) antiparallel β-strands followed by an amphipathic α-helix. Randomized residues of designed libraries are shown in blue and red, and were mutated with NNS and NHK codons, respectively. The position of the randomized loop, extended or not, is labeled "Loop 2". (**B**) Alignment of designed libraries. Secondary structure elements are indicated below the sequences. X represents all residues and Z all residues except Gly, Cys, Arg and Trp.
doi:10.1371/journal.pone.0097438.g001

Production and Purification of Proteins

HEWL was obtained from a commercial source (Sigma-Aldrich). CelD glycosidase (residues 34 to 577) was cloned into the pQE80 vector, which introduced a TEV cleavage site and a His-tag at the N-terminal of the protein. CelD was expressed in *E. coli* M15pREP4 (Qiagen). Affitins previously selected as in the "Analysis of selected pools and isolated clones" section, were expressed on a large scale. All cultures were grown to reach an OD_{600} of 1.2 in 2xYT and protein expression was induced with 0.5 mM IPTG for 16 h at 37°C for CelD and at 30°C for Affitins. Cells were pelleted by centrifugation at 4000 g for 15 min and resuspended in lysis buffer (50 mM $NaPO_4$ pH 7.5, 500 mM NaCl, 20 mM imidazole, 1 mg/ml HEWL, except for the anti-HEWL binder where HEWL was omitted) and frozen at −80°C. Pellets were thawed, sonicated and centrifuged at 18,000 g for 45 min. After purification by immobilized metal ion affinity chromatography (IMAC), using Chelating Sepharose Fast Flow resin charged with Ni^{2+} (GE Healthcare), proteins were injected into a Superdex75 16/60 column (GE Healthcare) equilibrated with 40 mM Tris-HCl pH 7.7 for CelD or 25 mM Tris-HCl pH

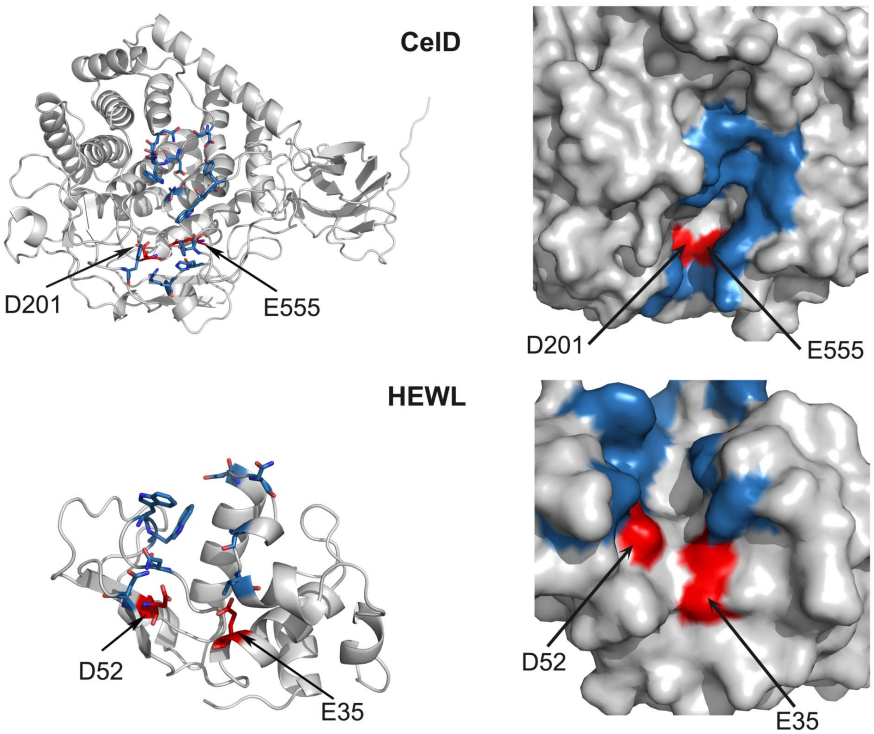


Figure 2. Crystal structures of CelD and HEWL glycosidases. Glycosidases (PDB codes: 4CJ1 and 4CJ2, respectively) are colored in gray with catalytic clefts in blue and catalytic residues in red. The right panel is a zoom view of active sites to show how catalytic residues are buried in CelD and less in HEWL.
doi:10.1371/journal.pone.0097438.g002

8.0, 500 mM NaCl for Affitins. The His-tag of purified CelD was cleaved with TEV protease. Uncleaved proteins, His-tag peptides and TEV proteases were removed by a second IMAC purification step. Purified proteins were quantified spectrophotometrically at 280 nm according to their molar extinction coefficients. Finally, for ITC and DSC analyses, proteins were desalted to PBS by using a HiPrep 26/10 desalting column (GE Healthcare).

Thermostability Measurements

DSC experiments were carried out in PBS, in a VP-DSC instrument (Microcal, Northampton, MA) and data analyzed with the software supplied with the equipment. The temperature was increased by 1°C per min from 30 to 120°C, and proteins were added at concentrations of 195, 217 and 300 μM for E12, H3 and H4 Affitins, respectively.

Isothermal Titration Microcalorimetry

ITC experiments were conducted using a VP-ITC instrument (Microcal, Northampton, MA). Injections of 10 μl of the different Affitins were added from a computer–controlled microsyringe at intervals of 460 s into the sample solution containing CelD or HEWL under constant stirring (400 rpm) at 25°C. The concentrations used for the experiments were 9.5 μM for CelD; 6.5 μM for HEWL and 195, 217 and 157 μM for E12, H3 and H4 Affitins, respectively. Titrations were carried out in PBS buffer. Data analysis was performed using Origin7 (Microcal), after subtraction of a manually-corrected baseline generated using constant heat values at the end of titration. Binding isotherms were fitted to a simple 1:1 Langmuir model. The same experiments were carried-out at 60°C for H3-CelD and E12-CelD.

Enzymatic Inhibition Assays

CelD activity was determined by a colorimetric assay using ρ-nitrophenyl-β-D-cellobioside (p-NPC, Sigma-Aldrich) as substrate; 500 nM of CelD was incubated with 0.5 mM of substrate for 1 h at room temperature or 60°C in PBS. The color change was measured spectrophotometrically at 415 nm and the final value corresponded to 100% of relative activity. HEWL activity was determined by monitoring the change in turbidity at 450 nm of a suspension of *M. lysodeikticus* bacteria (Sigma-Aldrich) in 100 mM potassium phosphate buffer, pH 7.0, as reported in [11]. Briefly, 400 μg/ml of cells was incubated with 20 nM of HEWL at RT for 1 h and the absorbance was measured. Enzymatic inhibition assays were carried out with different molar ratios of enzyme: Affitin (1:1, 1:2, 1:5 and 1:10). For the determination of the K_i values of anti-CelD Affitins, inhibition was carried out in PBS at 25°C, in the presence of 200 nM of CelD for 35 min. The substrate concentration (p-NPC) was, 5, 3, 2, 1, 0.5, 0.2, 0.05 and 0.02 mM. The inhibitor concentration (E12 or H3) was 0, 20, 50, 100 and 200 nM. Experiments were carried out in triplicate and fitted to a competitive inhibition model for anti-CelD or "One site–Fit K_i" for anti-HEWL Affitins using GraphPad Prism software (GraphPad Software).

Crystallization of Complexes

Anti-CelD Affitins and CelD were mixed in a 2:1 molar ratio to obtain a final concentration of 20 mg/ml for CelD in 25 mM Tris-HCl pH 8.0 and 100 mM NaCl. Affitin H4 and HEWL were mixed in a 1:1 molar ratio and the complex was purified by gel filtration chromatography with a Superdex75 16/60 column, equilibrated with the same buffer. The purified complex was concentrated to 80 mg/ml before setup crystallization trials. A crystallization screening was performed by mixing the complex

Table 1. Binding affinity, thermal stability parameters and inhibition constants of anti-glycosidase Affitins.

Target, Affitin	K_D, nM[a]	ΔG, kcal/mol	ΔH, kcal/mol	$T\Delta S$, kcal/mol	n	K_i, nM	Tm, °C[b]
HEWL, H4	11±2.8	−10.84	−16.77	−5.93	0.92	45±2	67.4
CelD, H3	48±10	−9.96	−6.69	3.27	0.93	111±10	76.1
CelD, E12	98±36	−9.56	−7.87	1.69	0.90	95±11	81.1

[a]Affinity obtained by ITC analysis at 25°C;
[b]Thermal melting obtained by DCS analysis. Errors shown derived from fitting to a 1:1 binding model for the K_D, and from a competitive inhibition or "One site–Fit" model for the K_i.
doi:10.1371/journal.pone.0097438.t001

```
                                        Loop 2
Sac7d  MVKVKFKYKG EEKEVDTSKI KKVWRVG--- -KMVSFTYDD NGKTGRGAVS EKDAPKELLD MLARAE
   A3  ......A... .......... .....YNTKL GYS.L..... .......I.R .S........ ......
  A12  ......R... .......... .....KLSKM GMY.F..... .......T.R .T........ ......
  E11  ......VKN. .......... .....HLSKM GMY.F..... .......T.T .T........ ......
  B12  ......LAA. .......... .....MLSKL GFY.M..... .......H.R .S........ ......
  B11  ......TEL. .......... .....NLSKF GML.L..... .......L.R .P........ ......
  A11  ......MLH. .......... .....YLSKL GVI.Q..... .......T.H .I........ ......
   E3  ......ALH. .......... .....FLSKM GTK.I..... .......M.. .D........ ......
   E4  ......V.Q. .......... .....YLAKW GNI.T..... .......W.N .Y........ ......
   H8  ......A... .K........ .....FLSKM GSL..... .......W.Y .N........ ......
  B10  ......M... .......... .....MLTKH GVL.L..... .......Y.A .N........ ......
  C12  ......D.P. K......... .....MLTKH GVL.L..... .......Y.H .......... ......
   D6  ......G... .......... .....MLTKY GHI.Q..... .......Y.Q .H........ ......
  E12  ......VSS. .......... .....NLTKY GTI.Q..... .......Y.R .L........ ......
  F12  ......GFT. .......... .....ALTKL GHL.M..... .......M.A .Q........ ......
   H3  ......HQI. .......... .....INTRL GMR.A..... .......M.P .......... ......
   A7  ......QVY. .......... .....TNLYK SVK.T..... .......R.N AQ........ ......
```

Figure 3. Sequence of anti-CelD Affitins. Alignment of the sixteen clones selected by ribosome display against the CelD enzyme. The conserved motif (Leu-Ser/Thr-Lys) inside the randomized and extended β-hairpin 2 is labeled in bold letters. Affitins studied in this work are underlined (E12 and H3).
doi:10.1371/journal.pone.0097438.g003

with 480 different buffers (1:1) at 19°C using the hanging-drop vapor-diffusion method. The crystallization buffer for the HEWL-H4 complex was 20% PEG 8000 (w/v), 100 mM CAPS pH 10.5 and 200 mM NaCl. For anti-CelD Affitins, it was 100 mM HEPES pH 7.5, 10.4% PEG 8000 (w/v), and 500 mM calcium acetate. Crystals were frozen in 20% glycerol diluted with the crystallization buffer.

Diffraction Data Collection

X-ray diffraction data for HEWL and CelD complexes were collected at the European Synchrotron Radiation Facility (ESRF) beamlines ID14–4 and ID23–2, respectively. Data reduction and scaling were performed with XDS [27] and Aimless [28], respectively.

Structure Determination, Model Building and Refinement

Crystal structures of HEWL and CelD in complex with their respective Affitins were solved by molecular replacement using Phaser [29]. Partial molecular replacement solutions using either HEWL (PDB code, 1GWD) or CelD (PDB code, 1CLC) as search models displayed extra electron density readily interpretable as the Affitin chain, which was manually traced. The structures were refined with Buster [30] and alternating rounds of model rebuilding with Coot [31]. All models were subjected to a last round of anisotropic B-factor refinement with Refmac [32] before MolProbity [33] validation. All structural representations were prepared with Pymol [34]. Protein-protein interaction parameters were calculated using the PISA server (www.ebi.ac.uk/msd-srv/prot_int/pistart.html) and LIGPLOT [35]. Shape complementarity analysis was performed with the SC program included in the CCP4 suite using default settings [36].

Accession Numbers

The atomic coordinates and structure factors have been deposited in the Protein Data Bank with the following accession codes: 4CJ0 (CelD-E12), 4CJ1 (CelD-H3) and 4CJ2 (HEWL-H4).

Results

Library Designs

Endo-glycosidases have cleft-shaped active sites and it is well known that loops can penetrate clefts. The short loop connecting β3-β4 strands (hereafter called "loop 2") of Sac7d was demonstrated to participate in the recognition of human immunoglobulin in a previously isolated anti-IgG Affitin [23]. Thus, we investigated if an artificially-extended loop 2, with an additional four residues between Gly27 and Lys28, could mimic this binding mode (libraries L3 and L4, Figure 1A) and could be helpful for efficient enzymatic inhibition. For example, CelD has deeply buried catalytic residues (Figure 2).

Selection of Anti-glycosidase Affitins

Two pools of libraries were constituted including the presence of the short (L1+L2) or long (L3+L4) loop 2, and selections were then performed in parallel by ribosome display using immobilized CelD as a target protein. For the L3+L4 selection, an ELISA after the fourth round indicated the expected enrichment in specific CelD binders. Two more rounds were performed for the L1+L2 selection without detectable enrichment.

Characterization of Anti-glycosidase Binders

Sequence analysis expression and purification of selected binders. Ninety-four randomly picked individual clones were screened by ELISA. Sixteen showed significant and specific CelD binding and were sequenced (Figure 3). Sequences originating from both libraries used for this selection (L3+L4) were identified. The 16 clones represented a variety of sequences. The motif Leu-Thr/Ser-Lys inside the extended loop was conserved, except in Affitin H3, where Leu was changed to Asn and Lys to Arg, although the latter implies a conservation of positive charge at this position. The other randomized positions did not show a conserved sequence. These data suggest that the extended loop might contribute significantly to the binding and highlight the probable importance of a positively charged residue.

With the aim of comparing potentially different modes of binding, we also included Affitin H4, which was selected against HEWL from the L1 library [20], and does not display obvious sequence similarity with anti-CelD Affitins (Figure 4A). Affitins with substantial sequence differences (E12 and H3) and Affitin H4 were produced in *E. coli* and purified to homogeneity (Figure 4B–C). These Affitins were predominantly eluted in size-exclusion chromatography at the volume corresponding to monomers. The production yields were up to 40 mg/L of shake flask culture (E12 and H3), and higher for H4 (90 mg/L culture).

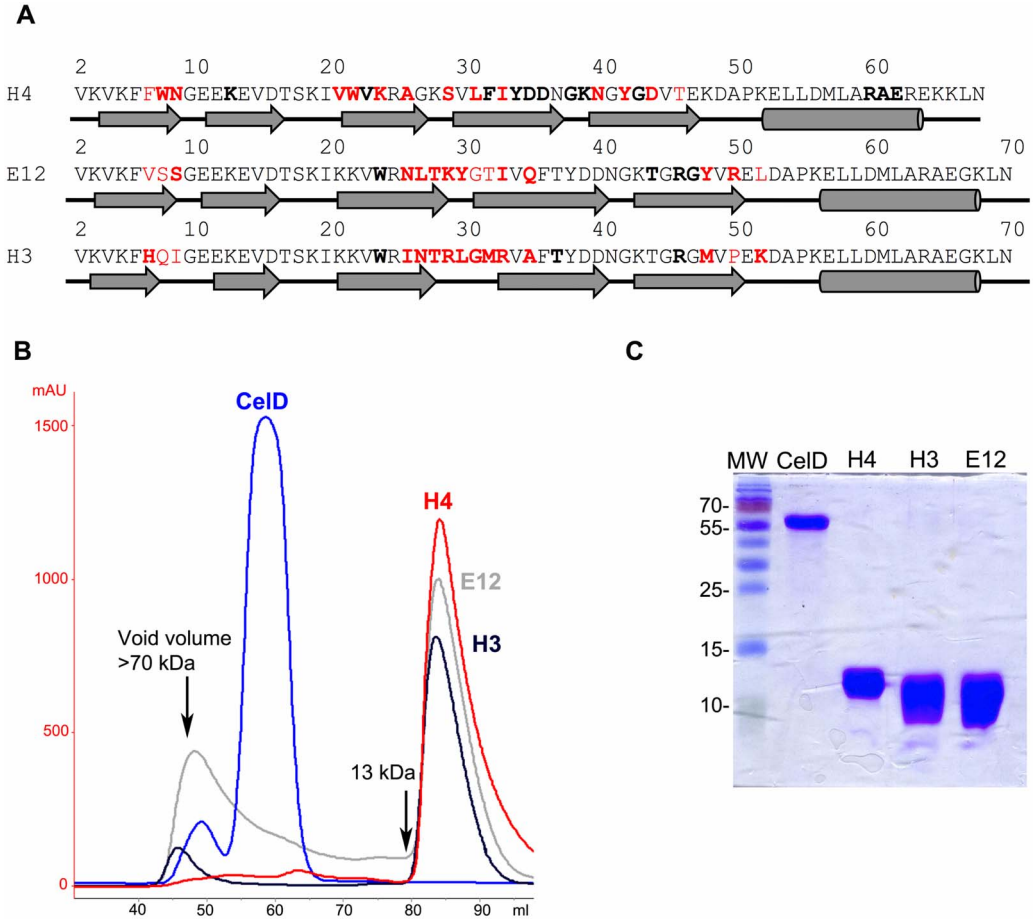


Figure 4. Sequences and production of anti-glycosidase Affitins. (**A**) Secondary structure elements according to crystallographic structure are shown below the sequences. Residues that were randomized are labeled in red. Residues that are involved in interaction with less than 10% of the buried surface appear in bold letters. (**B**) Size-exclusion purification of CelD (blue) and Affitin H4 (red), E12 (gray) and H3 (black) using a Superdex75 16/60 column. Arrows show the molecular weight obtained at the defined retention volumes. (**C**) SDS-PAGE 15% showing the final purity of CelD and the Affitins. Molecular weights are indicated in kDa.
doi:10.1371/journal.pone.0097438.g004

Binding properties of anti-glycosidase variants. The specificity of purified proteins was tested by ELISA analysis using the targets, streptavidin and bovine serum albumin (BSA) (Figure 5A). Affitins bound exclusively to their corresponding targets and not to unrelated proteins.

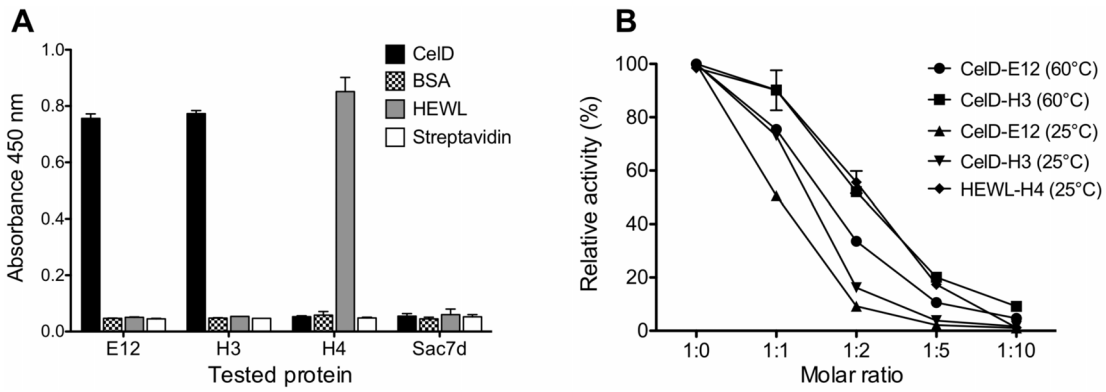


Figure 5. Biochemical properties of binders. (**A**) The interaction of E12, H3 and H4 Affitins (1 μM) was assayed by ELISA with immobilized CelD, streptavidin, BSA and HEWL. Sac7d wild-type was used as the negative control of binding at the same molar concentration. (**B**) Activity percentage of the thermophilic CelD glycosidase at 60°C and 25°C and of HEWL at 25°C. Different molar ratios (1:1, 1:2, 1:5 and 1:10) of Affitins were used as inhibitors.
doi:10.1371/journal.pone.0097438.g005

Table 2. Data collection and refinement statistics.

	HEWL-H4	CelD-E12	CelD-H3
Data collection:			
Resolution range (Å)	43.04–1.5 (1.53–1.5)	48.71–1.1 (1.12–1.1)	46.79–1.63 (1.66–1.63)
Space group	$P2_1$	$P4_3$	$P2_12_12_1$
Unit cell			
a, b, c (Å)	37.89, 62.82, 87.11	87.63, 87.63, 97.42	74.42, 97.73, 106.58
α, β, γ (°)	90, 98.7, 90	90, 90, 90	90, 90, 90
R-sym	0.036 (0.347)	0.109 (0.751)	0.053 (0.359)
R-meas	0.051 (0.491)	0.134 (0.916)	0.069 (0.473)
No. of unique reflections	62318 (3035)	292047 (14154)	96777 (4638)
I/σ(I)	16.9 (2.8)	10.5 (2.9)	15.4 (3.2)
Completeness (%)	96.3 (94.3)	98.4 (96.3)	99.7 (96.9)
Multiplicity	3.5 (3.4)	5.8 (5.7)	4.0 (3.9)
CC ½	0.998 (0.853)	0.997 (0.759)	0.998 (0.857)
Refinement:			
Resolution (Å)	1.50	1.10	1.63
No. of reflections	62318	277269	96678
R-factor/R-free	0.13/0.17	0.10/0.12	0.11/0.14
No. of atoms			
Macromolecules	3001	4711	4838
Ligands/ions	12	40	28
Water	358	866	781
RMS bonds (Å)	0.020	0.026	0.020
RMS angles (°)	1.884	2.030	1.814
B-factor (Å^2):			
Macromolecules	27.0	12.0	18.2
Ligands/ions	34.0	13.6	24.5
Water	36.3	27.8	32.8

Statistics for the highest-resolution shell are shown in parentheses.
doi:10.1371/journal.pone.0097438.t002

Isothermal titration microcalorimetry (ITC) binding analysis showed a stoichiometry close to 1, indicating a simple 1:1 binding mode of interaction for all binders (Figure 6A, Table 1). Affinity values determined for E12 and H3 CelD binders were in the nanomolar range (98 nM and 48 nM, respectively), while the affinity value measured for H4 (11 nM) was similar to a value obtained by surface plasmon resonance analysis [20]. Since Affitins and CelD enzyme [37] are thermostable (Figure 7, Table 1), we also determined K_D values of anti-CelD binders at 60°C, the optimal temperature for the activity of CelD; they were 176 and 157 nM for H3 and E12, respectively (Figure 6C). These results indicate that, although these Affitins were selected at 4°C, they showed an ability to interact with CelD with high affinity over a wide temperature range.

Thermodynamic parameters of the interactions were determined and indicated a favorable enthalpy for all cases and favorable entropy for H3 and E12 at 25°C (Table 1). Affitin H4 showed an enthalpy-driven reaction with a higher favorable binding enthalpy and unfavorable binding entropy, consistent with Affitins that bind through mutagenesis on the same surface [19].

Finally, the binding of anti-HEWL onto CelD and the binding of anti-CelD Affitins onto HEWL were tested by ITC. No cross-recognition could be observed (Figure 6B) further supporting the ELISA results.

Thermostability of anti-glycosidase Affitins. Thermal stabilities of E12 and H3 Affitins determined by differential scanning calorimetry (DSC) analysis were of 81.1°C and 76.1°C, respectively (Figure 7, Table 1). These results confirm that the extension of loop 2 is compatible with thermally stable Affitins. DSC scans were characteristic of cooperative unfolding, indicating that variants were well folded. However, both Affitins exhibited different behavior at high temperature. Affitin E12 showed no sign of protein aggregation after T_m was reached, while H3 showed an irreversible unfolding. H4 Affitin was also thermally stable and showed a primary T_m of 67.4°C with unfolding intermediates at higher temperatures.

Study of the Enzymatic Inhibition Properties of Affitins

The inhibitory properties of the purified Affitins were first analyzed at 25°C with different molar ratio of Affitins (Figure 5B). A ratio-dependent inhibition of activity was observed for the three binders. The K_i for the anti-CelD binders was determined as 111 nM for H3 and 95 nM for E12, while for the anti-HEWL binder a K_i value of 45 nM was obtained (Table 1). This latter value is similar to that observed for the HEWL-specific camel V_HH antibody (50 nM) [11,15]. The differences between the K_D and K_i values for Affitin H4 could result from experimental errors associated with the measurement of cell lysis for K_i determination.

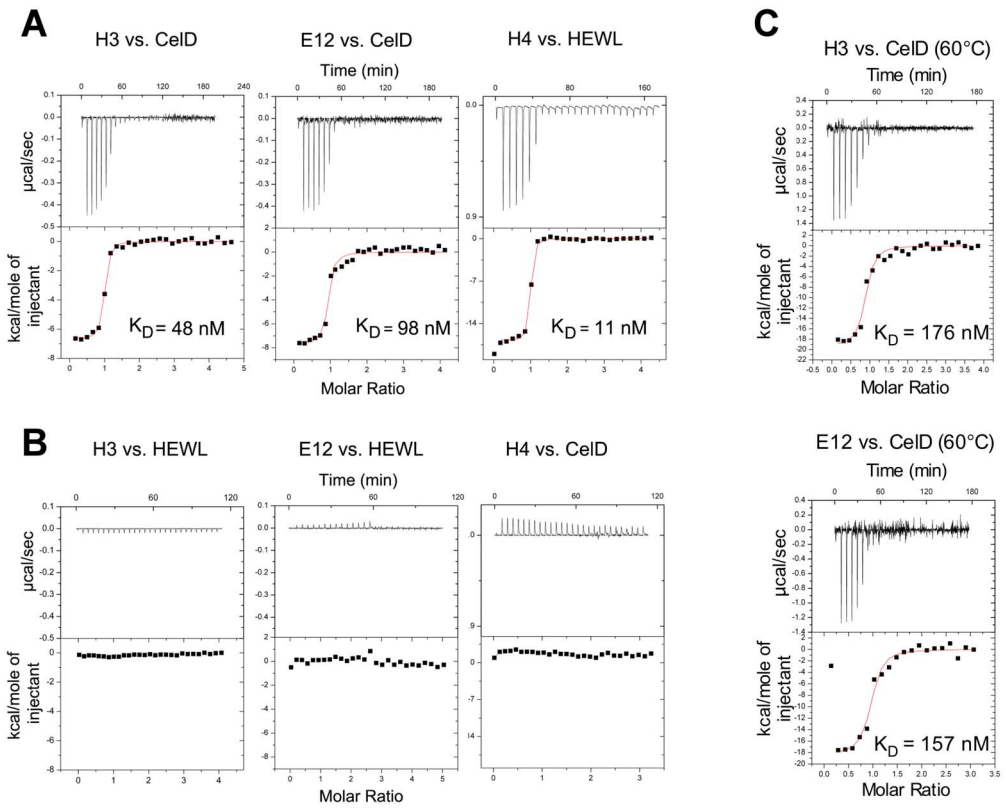


Figure 6. ITC analysis of anti-glycosidase binders. (**A**) ITC titrations at 25°C of Affitin H3 with CelD, Affitin E12 with CelD and Affitin H4 with HEWL. (**B**) Cross-recognitions were tested at 25°C for Affitin H3 with HEWL, Affitin E12 with HEWL and Affitin H4 with CelD. (**C**) ITC titrations at 60°C of Affitins H3 and E12 with CelD. The top panel for ITC shows data obtained from injections of Affitins while the bottom panel shows the integrated curve showing experimental points (filled squares) and the best fit (red line).
doi:10.1371/journal.pone.0097438.g006

CelD is a thermophilic glycosidase from *Clostridium thermocellum* and its optimal temperature for catalysis is 60°C [38]. As H3 and E12 Affitins are thermostable, it was possible to show that their inhibition properties at 60°C (Figure 5B) were similar to those determined at 25°C.

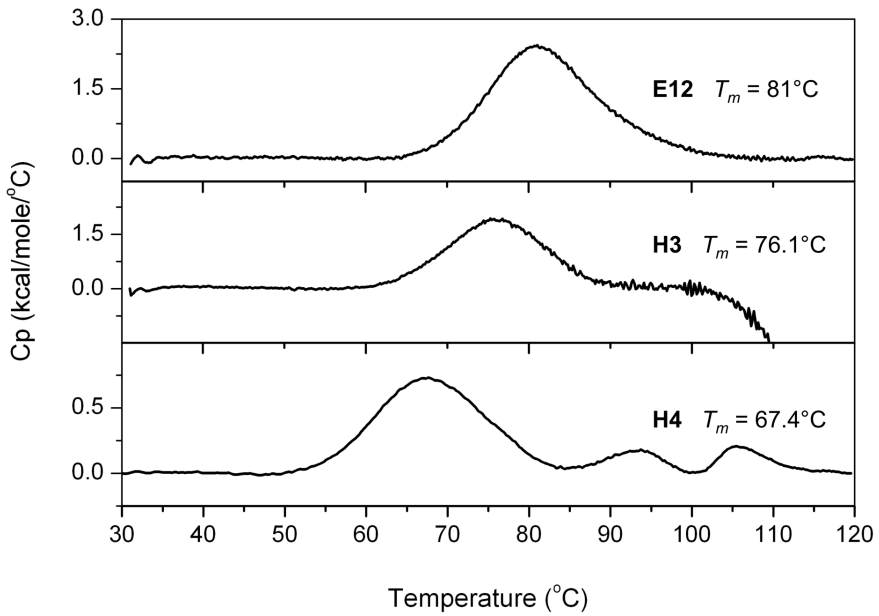


Figure 7. Thermal stabilities of anti-glycosidase binders. DSC curves of E12, H3 and H4 Affitins.
doi:10.1371/journal.pone.0097438.g007

Table 3. Interaction analysis of anti-glycosidase Affitins.

Affitin	Randomized region	H-bonds[a,b]	Salt-bridges[a]	Hydrophobic contacts[b]	Affitin BSA[a]	Complex BSA[a]
H4	Surface	11	2	10	838.7	1749.7
H3	Surface + loop	6	6	11	717.5	1317.3
E12	Surface + loop	5	2	7	707.1	1316.7

[a]Interaction contacts analyzed with the PISA server.
[b]Data obtained with Protein Interactor Calculator at 5 Å cutoff.
BSA: Buried surface area (Å^2), calculated with a water probe of 1.4 Å diameter.
doi:10.1371/journal.pone.0097438.t003

Crystal Structures of Affitin-enzyme Complexes

To analyze interactions at the atomic level, the crystal structures of the CelD-E12, CelD-H3 and HEWL-H4 complexes were determined at 1.1, 1.6 and 1.5 Å resolution, respectively (Figure 8). All complexes were crystallized in different crystal forms and the structures were solved by molecular replacement techniques using available enzyme structures as search models. Data collection and refinement statistics are presented in Table 2. Neither CelD nor HEWL structures underwent significant conformational changes upon Affitin binding, with RMSDs between Affitin-bound and ligand-free enzyme structures are 0.292 and 0.171 Å, respectively. Despite the large number of mutations and insertions in Sac7d, the overall fold was preserved in the three Affitins, *i.e.* an SH3-like five-stranded incomplete β-barrel capped by a C-terminal α-helix. As previously noticed with an anti-human IgG Affitin [23], the conserved β-barrel core did not show significant deviations when compared with the X-ray structure of wild-type Sac7d PDB code: 1AZP (RMSD <0.45 Å). Interestingly, we expected from our library designs a loop of 6 residues in length for H3 and E12 Affitins; however, it was partly structured in both cases by the extension of β3- and β4-strands (Figure 4). Calculated shape complementarity (Sc) values for each complex were 0.75, 0.72 and 0.76 for HEWL-H4, CelD-E12 and CelD-H3, respectively. These values are in agreement with those obtained by Lawrence and Colman [36] for protein/protein inhibitor interfaces (0.70–0.76), whereas for antibody/antigen interfaces Sc values are usually between 0.64–0.68.

Structural analysis of the interaction of the CelD-anti-CelD Affitins. E12 and H3 Affitins in complex with CelD displayed a similar interaction with an average buried surface area of 1317 Å^2 and an Affitin contribution of 707 and 717 Å^2, respectively (Figure 8, Table 3). In both cases, the Affitins bound the enzyme by inserting the protruding extended loop into the active site while the β-sheet 2 rested on the CelD surface. Additionally, the loop presented a charged residue that interacted *via* salt-bridges with those involved in the catalytic reaction.

The structure of the CelD-E12 complex suggested that Lys29 was the main residue responsible for binding and activity inhibition. It formed a salt-bridge through its nitrogen NZ with Glu555 of CelD, the residue that acts as a proton donor in

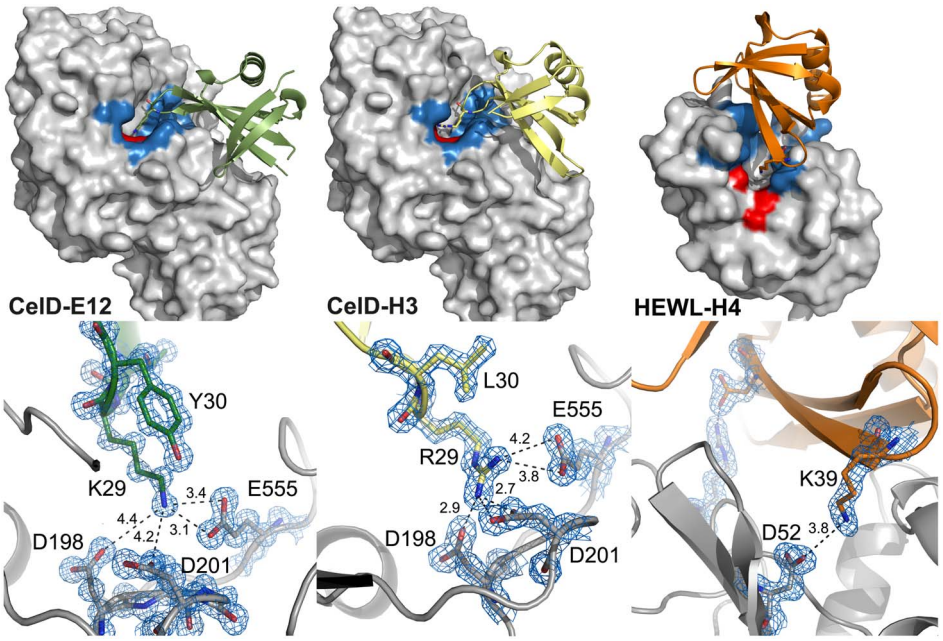


Figure 8. Crystal structures of anti-glycosidase Affitins in complex with their targets. Glycosidases are represented as gray surfaces with catalytic clefts colored in blue and catalytic residues in red. Affitins are represented in cartoons. The bottom panel shows a close-up view of the contacts and distances (Å) of the catalytic residues involved in salt-bridges and H-bonds (discontinuous lines). In blue, σA-weighted 2*mFobs - DFcalc* electron-density map contoured at 1.2 sigma for the HEWL-H4 complex, and at 2.0 sigma for CelD-H3 and CelD-E12 complexes. Residues and bond distances are indicated.
doi:10.1371/journal.pone.0097438.g008

catalysis [39]. Other key interactions involved E12 residues Gln35, Arg50 and Arg46 (non-randomized position), which were hydrogen-bonded to residues Tyr455, Tyr551 and Pro539 of CelD. These latter interactions stabilized the complex by positioning the β-sheet 2 over the CelD surface. In addition, due to Tyr551, which is positioned at the external part of the active site, the interaction with Arg50 limited access to the substrate even more. Finally, Lys29 also formed H-bonds *via* the main chain with Glu353 and Tyr354, which fix the artificial loop into the enzyme cavity.

For the CelD-H3 complex (Figure 8), the overall positioning of the Affitin onto CelD was similar. However, some contacts were different from those observed for the CelD-E12 complex. For example, Arg29 formed salt-bridges with other catalytically important residues (Asp198 and Asp201) [39,40].

Structural analysis of the interaction of the HEWL-anti-HEWL Affitins. The structure of Affitin H4 in complex with HEWL showed an unusual mechanism of inhibition (Figure 8). Unlike V_HH domains or anti-CelD Affitins, the randomized flat surface interacted with the enzyme by covering the catalytic site. A total buried surface area of 1750 $Å^2$ resulted from this interaction, to which Affitin H4 contributed 839 $Å^2$. This binding interface is larger than Affitins with the extended loop. There were seven residues from Affitin H4 involved in H-bonds, including the non-randomized Val23 and Gly38. Hydrophobic residues were found which spatially complement the interaction interface, especially residues Trp8, Trp23, and Tyr43 which are located inside the catalytic site, filling the cleft. These aromatic residues have a dual role forming intra- (Tyr43) and intermolecular (Trp8 and Trp23) H-bonds. Residues Lys39 and Asp44 formed two salt-bridges, which sealed the catalytic site by anchoring at the β5-strand ends of Affitin H4. The non-randomized Lys39 formed a salt-bridge with residue Asp52, which acts as a nucleophile to generate a glycosyl enzyme intermediate that is critical for HEWL activity. The interactions observed confirmed our previous results obtained by mutagenesis scanning of Affitin H4 [24].

Discussion

In this study, we demonstrate that specific and potent inhibitors of two glycosidases with at least two modes of binding can be derived from a unique scaffold protein. Catalytic residues are usually positioned inside a substrate-binding cleft or pocket on the enzyme's surface, and therefore molecules capable of binding deep inside these cavities or covering them can represent invaluable tools for glycosidase inhibition.

Small-molecule inhibitors are usually the preferred choice when targeting glycosidases, due to their pharmacological properties and because they can fit inside catalytic sites. About 1% of the human genome encodes for glycosyl processing enzymes [41], and among these 300 enzymes, 90 are glycosidases according to the CAZy database [42]. It is not ideal that small inhibitors mainly interact with catalytic residues often conserved among different glycosidases. Combining a high specificity and potency in one small molecule is thus difficult to achieve [6,8].

Proteinaceous inhibitors can bind to enzymes *via* a large surface area and are not limited to cavities. This enables them to interact with residues from non-conserved regions on the target, making this class of inhibitors potentially more specific. Artificially-generated inhibitors based on protein scaffolds are attractive since their properties, such as molecular weight, stability, lack of disulfide bridges or ease of production, can be chosen. In order for this approach to be generalized with minimal development effort, it is crucial that the same scaffold can bind to the different cavity shapes found in enzymes. With the design of several libraries and exploiting the high plasticity of the Sac7d scaffold, we were able to program different modes of binding in Affitins [23]. Here, we randomized a surface on Sac7d and extended the loop 2 with the aim of gaining loop flexibility and a potential to bind clefts. Using these different libraries, we obtained thermally stable binders with high affinity in the nanomolar range and specificity for thermostable CelD and for HEWL.

All three Affitins were shown to be inhibitors of two evolutionary distant endo-glycosidases, which both hydrolyze the O-glycosyl bond and have cleft-shaped catalytic sites but use two different enzymatic mechanisms. These anti-glycosidase Affitins have a K_i in the nanomolar range, which makes them comparable to the few best glycosidase inhibitors available that have a K_i $\sim 10^{-9}$ to 10^{-8} M [25]. Thus, we have obtained potent inhibitors with no cross-recognitions as shown by ELISA and ITC analysis. These Affitins are efficient inhibitors even at high temperatures (at least 60°C) although selected at 4°C. These could be useful as basic research tools to study *in vivo* biological events in thermophilic micro-organisms.

We have solved the crystal structure of the different complexes at high resolution, which shows that the recognized epitope is located in the catalytic cleft for both targets. The enzymatic inhibition properties are thus explained by hindrance of substrate access. Furthermore, the structures reveal that there is a direct interaction by H-bonds and salt-bridges with catalytically important residues in both enzymes, thereby locking the catalytic activity. The buried surfaces of the complexes (from 1317 $Å^2$ to 1749 $Å^2$) are comparable to natural protein-protein interactions [10]. Studies with other scaffolds have reported a modulation of the recognition by mutagenesis on their surface, on loop(s) or both [12,13,43]. Here, we present library designs providing Affitins using two modes of binding in an independent way, as shown by the structures of the complexes: by β-sheet 2 surface (Affitin H4), and a combination of β-sheet 2 surface and loop 2 (Affitins H3 and E12). E12 and H3 Affitins, which are derived from libraries with a longer and randomized loop 2, present a protruding convex region that penetrates the catalytic cleft of CelD, thereby validating our strategy to use an extended randomized loop. These structural data expand the possibilities of designing binding surfaces on Sac7d capable of recognizing different topographies in protein targets. They also provide useful hints for further inhibitor improvements, for example by randomizing residues that were kept constant in our library designs while they were identified in this work as interacting with targets. Importantly, no screen for enzymatic inhibition was performed to isolate the three Affitins that bind in two different catalytic sites. It remains to be seen if this is general but we believe this is not a fortuitous result, and suggests that Affitins have a propensity to bind where the curvature of the protein surface changes. In addition, the structures of Affitin-glycosidase complexes highlight that Affitins bind not only to catalytic-site residues but also to surrounding residues, contributing to their specificity. Variable domains of heavy-chain shark and camel anti-HEWL antibodies have been selected and structurally characterized [11,15,16,17]. Some of these were found to inhibit lysozyme activity by a mechanism similar to that reported here for anti-CelD Affitins. For instance, the CDR3 from a shark V-NAR was shown to be inserted into the HEWL active site and to engage in a salt-bridge interaction with the HEWL catalytic residue Asp52 [17].

For research or clinical purposes, it is important that the inhibitor does not interact with other glycosidases from the same organism of interest. We thus analyzed the alignment of sequences of all seventeen *C. thermocellum* (ATCC 27405) glycosidases from the GH9 CAZy family (including EC 3.2.1.4, EC 3.2.1.151, EC

3.2.1.91 enzymes) to which CelD belongs. We observed that among the residues of CelD interacting with Affitin E12 (Glu353, Tyr354, Val357, Tyr455, Trp538, Pro539, Tyr551, Glu555), three residues were not found in other glycosidases (Glu353, Val357, Trp538), while Pro539 was found in only three other glycosidases, suggesting a high specificity of E12. Furthermore, all these residues are outside or at the edge of the CelD catalytic site, confirming that E12 can recognize residues surrounding a catalytic site. We believe that such inhibitors might be a good starting point for the design of a new generation of low molecular weight drugs to modulate the activity of the most challenging targets in pharmaceutical research.

Protein-based therapeutics have been shown to be successful in clinical use and while monoclonal antibodies represent ~48% of these commercial recombinant proteins [44,45], there are also examples of non-human proteins, such as hirudin, which are used as therapeutics [46]. Given the difficulties related to antibody production, alternative scaffolds have recently been developed as binding molecules. Furthermore, several artificial affinity proteins with inhibition properties (for a review, see ref. [13]) derived from alternative scaffolds are undergoing clinical trials in the phases II/I [47] with the aim of using them as therapeutics. These include the Kunitz domain [48,49], Adnectin [50], and DARPins [51]. Monobodies are another source of binders that have been engineered to generate inhibitor molecules [52,53]. These alternative proteins present one or several attractive features, such as high-level expression in bacteria in soluble form, a simple monomeric structure, and stability toward denaturing agents and temperature. Although the performance of our Affitin-based class of inhibitors is yet to be evaluated *in vivo*, as demonstrated in the present and previous works, Affitins can be used as artificial binders and contain all these features with the additional property of resisting a wide pH range (usually from 0 to at least 10 and up to pH = 13). This combination of favorable properties and the resistance of Sac7d to harsh acidic conditions [23] may be interesting to inhibit targets within the digestive tract which are associated with pathologies such as α-glucosidase and diabetes type II [7].

We have previously described Affitins capable of inhibiting the type II secretion system (T2SS) in bacteria [19]. Here, we propose a strategy for generating potent glycosidase inhibitors with different modes of binding. We anticipate that Affitin-based inhibitors are not limited to glycosidases and may represent a generic method to obtain specific enzyme inhibitors with favorable properties interesting for research and clinical applications, and may provide an innovative approach for drug discovery.

Acknowledgments

We thank Ahmed Haouz of "Plateforme de Cristallogénèse et Diffraction des Rayons X" (Institut Pasteur). We also acknowledge the ESRF for provision of synchrotron radiation facilities and beamlines staff for their helpful assistance.

Author Contributions

Conceived and designed the experiments: AC SP AM GO BM PO PMA FP. Performed the experiments: AC SP AM GB. Analyzed the data: AC SP AM GO. Contributed reagents/materials/analysis tools: GB. Wrote the paper: AC SP PMA BM PO FP.

References

1. Bischoff H (1995) The mechanism of alpha-glucosidase inhibition in the management of diabetes. Clin Invest Med 18: 303–311.
2. Hruska KS, LaMarca ME, Scott CR, Sidransky E (2008) Gaucher disease: mutation and polymorphism spectrum in the glucocerebrosidase gene (GBA). Hum Mutat 29: 567–583.
3. Spearman MA, Ballon BC, Gerrard JM, Greenberg AH, Wright JA (1991) The inhibition of platelet aggregation of metastatic H-ras-transformed 10T1/2 fibroblasts with castanospermine, an N-linked glycoprotein processing inhibitor. Cancer Lett 60: 185–191.
4. Zhu Z, Zheng T, Homer RJ, Kim YK, Chen NY, et al. (2004) Acidic mammalian chitinase in asthmatic Th2 inflammation and IL-13 pathway activation. Science 304: 1678–1682.
5. Leeson PD, Springthorpe B (2007) The influence of drug-like concepts on decision-making in medicinal chemistry. Nat Rev Drug Discov 6: 881–890.
6. Gloster TM, Vocadlo DJ (2012) Developing inhibitors of glycan processing enzymes as tools for enabling glycobiology. Nat Chem Biol 8: 683–694.
7. Moorthy NS, Ramos MJ, Fernandes PA (2012) Studies on alpha-glucosidase inhibitors development: magic molecules for the treatment of carbohydrate mediated diseases. Mini Rev Med Chem 12: 713–720.
8. Cheng AC, Coleman RG, Smyth KT, Cao Q, Soulard P, et al. (2007) Structure-based maximal affinity model predicts small-molecule druggability. Nat Biotechnol 25: 71–75.
9. Carlson HA, Smith RD, Khazanov NA, Kirchhoff PD, Dunbar JB Jr., et al. (2008) Differences between high- and low-affinity complexes of enzymes and nonenzymes. J Med Chem 51: 6432–6441.
10. Jones S, Thornton JM (1996) Principles of protein-protein interactions. Proc Natl Acad Sci U S A 93: 13–20.
11. Transue TR, De Genst E, Ghahroudi MA, Wyns L, Muyldermans S (1998) Camel single-domain antibody inhibits enzyme by mimicking carbohydrate substrate. Proteins 32: 515–522.
12. Binz HK, Amstutz P, Pluckthun A (2005) Engineering novel binding proteins from nonimmunoglobulin domains. Nat Biotechnol 23: 1257–1268.
13. Gebauer M, Skerra A (2009) Engineered protein scaffolds as next-generation antibody therapeutics. Curr Opin Chem Biol 13: 245–255.
14. Stoop AA, Craik CS (2003) Engineering of a macromolecular scaffold to develop specific protease inhibitors. Nat Biotechnol 21: 1063–1068.
15. Desmyter A, Transue TR, Ghahroudi MA, Thi MH, Poortmans F, et al. (1996) Crystal structure of a camel single-domain VH antibody fragment in complex with lysozyme. Nat Struct Biol 3: 803–811.
16. De Genst E, Silence K, Decanniere K, Conrath K, Loris R, et al. (2006) Molecular basis for the preferential cleft recognition by dromedary heavy-chain antibodies. Proceedings of the National Academy of Sciences of the United States of America 103: 4586–4591.
17. Stanfield RL, Dooley H, Flajnik MF, Wilson IA (2004) Crystal structure of a shark single-domain antibody V region in complex with lysozyme. Science 305: 1770–1773.
18. Steemson JD, Baake M, Rakonjac J, Arcus VL, Liddament MT (2014) Tracking Molecular Recognition at the Atomic Level with a New Protein Scaffold Based on the OB-Fold. PLoS One 9: e86050.
19. Mouratou B, Schaeffer F, Guilvout I, Tello-Manigne D, Pugsley AP, et al. (2007) Remodeling a DNA-binding protein as a specific in vivo inhibitor of bacterial secretin PulD. Proc Natl Acad Sci U S A 104: 17983–17988.
20. Pecorari F, Alzari PM (2008) OB-fold used as scaffold for engineering new specific binders. Patent Publication Nos WO2008068637 (A3), EP2099817 (A2).
21. Krehenbrink M, Chami M, Guilvout I, Alzari PM, Pecorari F, et al. (2008) Artificial binding proteins (Affitins) as probes for conformational changes in secretin PulD. J Mol Biol 383: 1058–1068.
22. Buddelmeijer N, Krehenbrink M, Pecorari F, Pugsley AP (2009) Type II secretion system secretin PulD localizes in clusters in the Escherichia coli outer membrane. J Bacteriol 191: 161–168.
23. Behar G, Bellinzoni M, Maillasson M, Paillard-Laurance L, Alzari PM, et al. (2013) Tolerance of the archaeal Sac7d scaffold protein to alternative library designs: characterization of anti-immunoglobulin G Affitins. Protein Eng Des Sel 26: 267–275.
24. Miranda FF, Brient-Litzler E, Zidane N, Pecorari F, Bedouelle H (2011) Reagentless fluorescent biosensors from artificial families of antigen binding proteins. Biosens Bioelectron 26: 4184–4190.
25. Vasella A, Davies GJ, Bohm M (2002) Glycosidase mechanisms. Curr Opin Chem Biol 6: 619–629.
26. Mouratou B, Behar G, Paillard-Laurance L, Colinet S, Pecorari F (2012) Ribosome display for the selection of Sac7d scaffolds. Methods Mol Biol 805: 315–331.
27. Kabsch W (2010) Xds. Acta Crystallogr D Biol Crystallogr 66: 125–132.
28. Evans PR (2011) An introduction to data reduction: space-group determination, scaling and intensity statistics. Acta Crystallogr D Biol Crystallogr 67: 282–292.
29. McCoy AJ (2007) Solving structures of protein complexes by molecular replacement with Phaser. Acta Crystallogr D Biol Crystallogr 63: 32–41.
30. Bricogne G BE, Brandl M., Flensburg C., Keller P., Paciorek W., Roversi P SA, Smart O.S., Vonrhein C., Womack T.O (2011) BUSTER version 2.11.1. Cambridge, United Kingdom: Global Phasing Ltd.
31. Emsley P, Cowtan K (2004) Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr 60: 2126–2132.

32. Murshudov GN, Vagin AA, Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D Biol Crystallogr 53: 240–255.
33. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, et al. (2007) MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res 35: W375–383.
34. DeLano WL (2002) The PyMOL Molecular Graphics System.
35. Laskowski RA, Swindells MB (2011) LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. J Chem Inf Model 51: 2778–2786.
36. Lawrence MC, Colman PM (1993) Shape complementarity at protein/protein interfaces. Journal of Molecular Biology 234: 946–950.
37. Correa AC, Ortega CO, Obal GO, Alzari PA, Vincentelli RV, et al. (2014) Generation of a vector suite for protein solubility screening. Frontiers in Microbiology 5: 1–9.
38. Peng J, Wang W, Jiang Y, Liu M, Zhang H, et al. (2011) Enhanced soluble expression of a thermostable cellulase from Clostridium thermocellum in Escherichia coli. Curr Microbiol 63: 523–530.
39. Chauvaux S, Beguin P, Aubert JP (1992) Site-directed mutagenesis of essential carboxylic residues in Clostridium thermocellum endoglucanase CelD. J Biol Chem 267: 4472–4478.
40. Juy MA, Adolfo G.; Alzari, Pedro M.; Poljak, Roberta J.; Claeyssens, Marc; Béguin, Pierre; Aubert, Jean-Paul (1992) Three-dimensional structure of a thermostable bacterial cellulase. Nature 357: 89–91.
41. Davies GJ, Gloster TM, Henrissat B (2005) Recent structural insights into the expanding world of carbohydrate-active enzymes. Curr Opin Struct Biol 15: 637–645.
42. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, et al. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res 37: D233–238.
43. Koide A, Wojcik J, Gilbreth RN, Hoey RJ, Koide S (2012) Teaching an old scaffold new tricks: monobodies constructed using alternative surfaces of the FN3 scaffold. J Mol Biol 415: 393–405.
44. Dimitrov DS, Marks JD (2009) Therapeutic antibodies: current state and future trends–is a paradigm change coming soon? Methods Mol Biol 525: 1–27, xiii.
45. Dimitrov DS (2012) Therapeutic proteins. Methods Mol Biol 899: 1–26.
46. Investigators O- (1999) Effects of recombinant hirudin (lepirudin) compared with heparin on death, myocardial infarction, refractory angina, and revascularisation procedures in patients with acute myocardial ischaemia without ST elevation: a randomised trial.. Lancet 353: 429–438.
47. Wurch T, Pierre A, Depil S (2012) Novel protein scaffolds as emerging therapeutic proteins: from discovery to clinical proof-of-concept. Trends Biotechnol 30: 575–582.
48. Williams A, Baird LG (2003) DX-88 and HAE: a developmental perspective. Transfus Apher Sci 29: 255–258.
49. Attucci S, Gauthier A, Korkmaz B, Delepine P, Martino MF, et al. (2006) EPI-hNE4, a proteolysis-resistant inhibitor of human neutrophil elastase and potential anti-inflammatory drug for treating cystic fibrosis. J Pharmacol Exp Ther 318: 803–809.
50. Tolcher AW, Sweeney CJ, Papadopoulos K, Patnaik A, Chiorean EG, et al. (2011) Phase I and pharmacokinetic study of CT-322 (BMS-844203), a targeted Adnectin inhibitor of VEGFR-2 based on a domain of human fibronectin. Clin Cancer Res 17: 363–371.
51. Campochiaro PA, Channa R, Berger BB, Heier JS, Brown DM, et al. (2012) Treatment of Diabetic Macular Edema With a Designed Ankyrin Repeat Protein That Binds Vascular Endothelial Growth Factor: A Phase 1/2 Study. Am J Ophthalmol.
52. Wojcik J, Hantschel O, Grebien F, Kaupe I, Bennett KL, et al. (2010) A potent and highly specific FN3 monobody inhibitor of the Abl SH2 domain. Nat Struct Mol Biol 17: 519–527.
53. Gilbreth RN, Truong K, Madu I, Koide A, Wojcik JB, et al. (2011) Isoform-specific monobody inhibitors of small ubiquitin-related modifiers engineered using structure-guided library design. Proc Natl Acad Sci U S A 108: 7751–7756.

DISCUSIÓN Y CONCLUSIONES

El desarrollo de la tecnología del ADN recombinante, permitió la producción de las proteínas tanto en sistemas eucariotas como procariotas. De esta manera fue posible expresar y purificar proteínas recombinantes para su caracterización funcional a través de distintas tecnologías como difracción de rayos X, espectroscopia de resonancia magnética nuclear (NMR), calorimetría, entre otras. Además, permitió la modificación del gen de interés pudiéndose evaluar diferentes variantes, mutantes o generar quimeras que nunca se habían dado en la naturaleza. Sin embargo y a pesar de los importantes avances en el área, en muchos casos no es posible producir la proteína recombinante en forma soluble, homogénea y activa. A diferencia de lo que ocurre con la purificación de por ejemplo ácidos nucleicos, no existe al día de hoy un protocolo genérico que garantice la purificación en forma soluble de las PRs, requiriéndose en muchos casos una extensiva evaluación de condiciones de expresión (2, 3).

Los avances de miniaturización de los cultivos así como el desarrollo de las metodologías de HTS, han permitido realizar evaluaciones de cientos o incluso miles de condiciones de expresión de manera automática, disminuyendo el tiempo requerido y los costos (79, 288). Esto se ve reflejado en los consorcios de genómica estructural que están basados en el uso de estas metodologías HTS para la expresión y cristalización de miles de proteínas de forma automática. Por ejemplo, en el Consorcio de Genómica Estructural del Noreste (Northeast Structural Genomics Consortium), las etapas de clonado, evaluación de la expresión y purificación, se realizan de manera automática y en forma HTS. Cada semana más de 100 proteínas blanco son clonadas mediante técnicas como In-Fusion o LIC y su expresión evaluada, donde 50-75 constructos son expresados en fermentaciones preparativas (1-2lt). De esta manera, 30-40 blancos son purificados en cantidades de decenas de miligramos para su caracterización biofísica incluyendo NMR (espectroscopia de resonancia magnética nuclear) y/o cribado de cristalización (115). En 10 años, en este centro por ejemplo se clonaron más de 39000 genes, lográndose la expresión soluble en cantidades mayores a 0.5 mg de proteína pura para más de 6900 proteínas o dominios. Esto llevó al depósito de más de 650 estructuras cristalográficas y 500 estructuras por NMR (http://nesg.org/statistics.html).

A nivel de laboratorio, también es posible evaluar cientos de condiciones de expresión sin la necesidad de sistemas robóticos para las etapas de cribado y con costos accesibles. El equipamiento necesario para la realización de un cribado de expresión de PRs en un laboratorio estándar, es relativamente básico consistiendo principalmente en placas de 96 y 24 pocillos, pipetas multicanal, placas filtrantes de 96 pocillos, sistemas de vacío y resinas para las etapas de purificación (5, 78, 79). Sin embargo, para tener éxito en la producción soluble de

una proteína difícil de expresar, muchas veces es necesario evaluar diferentes proteínas de fusión así como también diferentes promotores, para lo cual se requiere el clonado del gen de interés en varios vectores. Realizar esto por métodos basados en enzimas de restricción puede ser una tarea complicada, principalmente cuando varios sitios de restricción están presentes y más aún si se está trabajando con varios genes en simultáneo. Esto puede ser un factor limitante si no se cuenta con un método de clonado eficiente e independiente de la secuencia del gen.

En este sentido, diversas estrategias se han desarrollado que facilitan la transferencia de un mismo fragmento de ADN a varios vectores en simultáneo y están basadas en recombinación o apareamiento simple hebra de ADN. Una metodología muy utilizada es Gateway (Invitrogen) basada en una recombinación sitio específica. Si bien ha mostrado ser una técnica muy eficiente, la recombinación se da entre sitios *att* diferentes de 21 pb (que pueden ser *attB, attP, attL* y *attR*) (289), por lo que esta metodología tiene tres limitantes importantes. Primero, el gen de interés solo se insertará en los mencionados sitios limitando la posibilidad de realizar inserciones alternativas en el vector. Segundo, debido a la necesidad de estos 21 pb es que la proteína expresada contendrá al menos 7 residuos adicionales en su extremo amino terminal lo que puede afectar estudios posteriores. Por último, para la reacción de recombinación se necesita de enzimas específicas que son costosas limitando su aplicación. Otra metodología desarrollada posteriormente a Gateway y que se basa en la recombinación homóloga es InFusion (290). Esta técnica tiene como ventaja que no requiere de los sitios *att*, por lo que se evita la inserción de residuos extra a la proteína a expresar. Para la misma se adicionan a los cebadores 15-20 pb que son complementarios al sito de inserción. Una vez amplificado el fragmento el mismo se incuba con el vector de destino linearizado (tras digestión enzimática), durante 30 minutos en presencia de la enzima InFusion, que genera hebras de ADN simple hebra y que luego hibridan debido a que hay complementariedad de bases (290, 291). Si bien esta metodología sortea una de las desventajas encontradas en Gateway, es necesario tratar al vector con enzimas de restricción para linearizarlo (limitando por ejemplo la inserción de un fragmento de ADN en sitios del vector que no tengan enzimas de restricción) y se requiere de una enzima específica para la generación de las hebras simples, aumentando los costos. Una estrategia similar a In-Fusion es conocida como LIC (del inglés, ligation independent cloning) (292). La idea general es similar a InFusion por lo que secuencias complementarias son adicionadas a los cebadores, pero la enzima utilizada para la generación de ADN simple hebra es la T4 polimerasa (293). Como para el caso de InFusion, también se necesita del vector lineal. Una nueva metodología, que permite clonar un fragmento de ADN en cualquier sitio en el

vector sin la necesidad de que este lineal, no adiciona secuencias extras en el gen y utiliza reactivos comunes de PCR, es el clonado libre de restricción (RF cloning) (4). En este método el ADN es amplificado con cebadores que poseen 25-30 nucleótidos complementarios con el sitio de inserción en el vector de destino. Luego de la amplificación, el producto es utilizado como megaprimer en una segunda reacción de PCR para amplificar el plásmido entero e insertando la secuencia deseada en el sitio seleccionado. Finalmente, el plásmido original se digiere con DpnI para eliminar el vector parental y se transforman células de *E. coli*. Recientemente se desarrolló una herramienta web que permite diseñar de manera fácil los cebadores para realizar clonado RF (http://www.rf-cloning.org) (294). Por lo tanto el clonado RF es una metodología muy versátil, simple, económica y fácil de implementar.

De esta forma varios grupos han desarrollado sus propias series de vectores para la expresión de proteínas recombinantes que permiten clonar un mismo fragmento de ADN en forma paralela. Por ejemplo, el grupo de Bottomley generó una serie de 4 vectores donde el mismo fragmento de ADN puede ser clonado mediante LIC. Con los mismos, la proteína puede expresarse con un HisTag N-terminal o alguna de las fusiones MBP, GST y NusA. Estos vectores contienen un promotor T7 y un sitio TEV para el corte de la fusión (295). También para clonado por LIC, se generaron 5 vectores, que permiten evaluar 3 promotores (T7, tac y CspA), diferentes tags de purificación (HisTag y strepTag II), así como las proteínas de fusión GST y TF (del inglés Trigger Factor). Además se generó un vector con el mismo sitio de inserción que para los 5 anteriores pero para expresión en células de insecto (296). Por otro lado, el grupo de Peleg, generó una serie de 5 vectores para la evaluación de la expresión bajo el control del promotor T7, de fusiones con MBP, GST, DsbC, DsbA y GB1 (dominio β1 de la proteína streptococal G) que al contener el mismo sitio de inserción son apropiados para el clonado RF, por lo que con 2 cebadores es posible clonar el gen de interés de manera simultánea en los 5 vectores (4).

Enmarcado en este trabajo de Tesis, hemos generado en nuestro laboratorio una serie de vectores optimizados para la expresión de PRs en *E. coli*. Esta "suite" de vectores permite el clonado en paralelo de los genes de estudio y genera una herramienta disponible en el marco de la unidad de servicio de PRs para la expresión y evaluación de diferentes condiciones de expresión. Como primer paso seleccionamos la metodología de clonado RF ya que como se mencionó anteriormente es simple, versátil y puede utilizarse tanto para la generación de los vectores como para el posterior clonado de los genes de interés en estos. Como variables a introducir en los vectores decidimos probar diferentes promotores y distintas proteínas de

fusión en combinación, permitiendo una evaluación sistemática del efecto de dichos parámetros.

Como elementos constantes se optó por un único marcador de selección (ampicilina), de manera de facilitar las etapas de cultivo, y un sitio de reconocimiento para el corte proteolítico por la proteasa TEV. Esta proteasa tiene como ventajas la alta especificidad, posibilidad de realizar el corte a bajas temperaturas (4˚C) y su expresión en el laboratorio con altos rendimientos, reduciendo los costos (112). Además la ultima glicina presente en el sitio de reconocimiento, puede ser sustituida por cualquier otro residuo, excepto prolina, pero a expensas de eficiencia de corte, por lo que podría liberarse un producto con un extremo amino terminal nativo si fuese necesario (297). Todos los vectores contienen además un HisTag amino terminal y un strepTag II carboxilo terminal. El HisTag es el tag de afinidad más comúnmente utilizado ya que es de pequeño tamaño, tiene condiciones de elusión suaves y se han desarrollado protocolos automáticos para la purificación con este tag a pequeña y gran escala (120, 298, 299). Sin embargo como hemos mencionado anteriormente, cuando se trabaja con proteínas que se expresan con muy bajos rendimientos, muchos contaminantes son co-purificados luego de la cromatografía de afinidad IMAC (84, 85). Es por este motivo que la incorporación de un segundo tag como el strepTag II en nuestra serie ofrece varias ventajas. Por un lado es más específico que el HisTag, por lo que se evitan los contaminantes que co-purifican en IMAC (88, 89). Por otro lado, al no verse afectado por los agentes quelantes de la bacteria, los rendimientos en las etapas de escalado no son afectados negativamente. Finalmente, al acoplar un primer paso de purificación por IMAC con un segundo paso de purificación con el strepTag II, se garantiza una proteína recombinante con ambos extremos intactos.

Como promotores seleccionamos dos que son los más utilizados al momento, el promotor T7 proveniente de la serie de vectores pET (Novagen) y el promotor T5, proveniente de la serie de vectores pQE (QIAGEN). Mientras que el promotor T7 tiene una tasa de transcripción muy alta, el T5 tiene un control de la expresión más a justado a tasas de transcripción menores (30).

Otra variable que introdujimos en los vectores es el uso de diferentes proteínas de fusión. Las mismas han mostrado potenciar la solubilidad de muchas proteínas blanco que de otra forma se expresaban insolubles o con rendimientos muy bajos (92). Si bien aún no se ha encontrado una proteína de fusión que garantice la expresión soluble del blanco o que funcione mejor que el resto de las proteínas de fusión, sí se han visto ciertas tendencias. Por ejemplo en un estudio comparativo se observó que MBP permitía la obtención de la mayor cantidad de proteínas en

estado soluble (50 proteínas), aunque la mitad precipitaban luego del corte proteolítico. Para el caso de NusA, si bien también permitía la expresión de varias proteínas (39 proteínas) sólo el 15% permanecía soluble luego del corte proteolítico. Por el contrario, en las fusiones con Trx, todas las fusiones solubles (24 proteínas), permanecieron solubles incluso luego del corte proteolítico (78). En otros estudios, se demostró que la fusión con DsbC era la más eficiente en la expresión soluble de proteínas pequeñas ricas en puentes S-S (67), mientras que la fusión con SUMO podía ser más efectiva para la expresión de una proteína blanco comparado con MBP, GST, Trx y NusA (103).

Además, nuevas proteínas potenciadoras de la solubilidad han sido evaluadas por diversos grupos como ser por ejemplo el mutante de la proteína Ocr del bacteriófago T7 (Mocr, 13.8 kDa) (300). Utilizando este tag, se logró la expresión soluble de 4 proteínas blanco con niveles similares a los obtenidos con fusiones con MBP (300). Otro grupo utilizó la quinasa del bacteriófago T7 (T7PK, 30 kDa) y la chaperona skp (17 kDa) de *E. coli* para la expresión de 4 proteínas humanas. Mientras que las proteínas blanco permanecían insolubles cuando eran expresadas sin fusión, se lograron obtener de forma soluble al realizar las fusiones con las proteínas propuestas (301). Finalmente, se realizaron fusiones con la proteína de *E. coli* EspA (25 kDa). Al fusionar 6 proteínas al C-terminal de EspA, se encontró un aumento en la expresión para todas las fusiones (302).

Teniendo en cuenta estos estudios, seleccionamos como proteínas de fusión para incluir en nuestra serie de vectores a SUMO, Trx, DsbC y MBP. Además, incluimos una nueva proteína como proteína potenciadora de la solubilidad. Esta corresponde a la versión truncada de CelD (CelDnc) que, como mencionamos anteriormente, tiene propiedades de solubilidad y termoestabilidad atractivas para su uso como proteína de fusión. De esta forma se generaron exitosamente mediante clonado RF, una serie de 12 vectores donde el mismo producto de PCR puede ser clonado en todos ellos, utilizando solo dos cebadores. La expresión en los mismos permite evaluar el efecto de dos promotores (T5 o T7) en combinación con 5 proteínas de fusión (SUMO, Trx, DsbC, MBP y CelDnc) o solo con el HisTag.

Para la evaluación del correcto funcionamiento de los 12 vectores generados, se clonó una proteína que se expresa de manera soluble como ser GFP. Además, para probar la efectividad de la serie de vectores, se clonaron 2 proteínas cuya expresión no se había logrado de forma soluble. Estas correspondían a la proteína de *M. smegmatis* DprE1 y la proteína de *L. major* MPK4. El clonado en la serie de vectores se efectuó con alta eficiencia, ya que >80% de los clones obtenidos fueron positivos. Si bien este porcentaje es alto, todavía podría mejorarse

insertando la secuencia de un gen tóxico que inhiba el crecimiento de *E. coli*. Un ejemplo es el sistema toxina-antitoxina CcdB, el cual es capaz de inhibir la ADN girasa y es neutralizado por el gen ccdA, actualmente utilizado en el sistema Gateway (303)

Con las construcciones generadas para los genes GFP y DprE1, se realizó un cribado de expresión de forma manual y adaptable a laboratorios con equipamientos básicos, que permite la evaluación de 48 condiciones de expresión en 4 días incluyendo las etapas de: inoculación, inducción, lisis, purificación por IMAC, corte con TEV y visualización en SDS-PAGE 96x. En este caso se evaluaron 24 condiciones para ambos genes siendo 12 vectores a dos temperaturas de inducción diferentes (17 y 37˚C). Este protocolo es adaptable a la evaluación de diferentes cepas de expresión, así como también distintos medios de cultivo.

Tras la expresión del gen de GFP, pudimos observar que en todos los casos se logró expresar y purificar por IMAC las diferentes fusiones donde las fracciones purificadas tenían un intenso color verde, confirmando no solo la solubilidad sino también el correcto plegamiento de GFP. Más aún fue posible realizar el corte de las fusiones luego del tratamiento con TEV. Estos resultados muestran que los 12 vectores funcionan correctamente en las etapas de clonado y de expresión, donde los módulos de purificación (HisTag) y corte con TEV quedan correctamente expresados y accesibles.

Al analizar los resultados de expresión para el gen DprE1, se confirmó que no era posible expresarlo de forma soluble sólo con el HisTag. Por el contrario, se obtuvieron altos rendimientos de expresión con las diferentes fusiones. Más aún, éstas eran sensibles al corte con TEV, obteniéndose una banda correspondiente al producto DprE1 sin fusión (51 kDa). Estos resultados muestran la utilidad del uso de las proteínas de fusión para la expresión de proteínas recombinantes incorporadas en nuestro sistema. Al analizar las cantidades obtenidas para las distintas fusiones se determinó que la fusión con CelDnc producía rendimientos similares o incluso mayores a los obtenidos con proteínas de fusión ya conocidas como ser MBP, SUMO y Trx. Dado que la expresión soluble de las proteínas blanco sometidas a HTS deben ser confirmadas posteriormente en cultivos a mayor escala, realizamos un escalado (1lt en matraz de 5 lts) con la fusión CelD-DprE1. Nuestros resultados muestran que luego del corte con TEV y purificación por SEC, DprE1 se comporta como un monómero obteniéndose un rendimiento final de 7 mg/ml. Más aún, se observó que DprE1 podía unir FAD indicando un correcto plegamiento de la proteína de interés. Estos datos sugieren que CelDnc es capaz de aumentar y asistir en la expresión y correcto plegamiento de proteínas fusionadas a ella, en este caso DprE1. Esta primera parte del trabajo de Tesis sugiere entonces a CelDnc como una

herramienta potencialmente útil para lograr la expresión soluble de otras proteínas de difícil expresión en *E. coli*.

El cribado de expresión automático para la proteína MPK4, mostró que la fusión con DsbC es la única condición en donde la fusión podía expresarse de forma soluble y es sensible al corte por TEV. Tras el escalado, se lograron obtener buenos rendimientos de la fusión completa (6 mg/lt), sin embargo tras el corte con TEV, la gran mayoría del producto generado precipitaba. Al analizar el estado oligomérico de la fusión, la misma no mostró ser un agregado soluble sino formar un decámero homogéneo que podría ser utilizado igualmente en ensayos de cristalización.

Dentro de las perspectivas para mejorar la serie de vectores de expresión, proponemos la inserción del gen que codifica CcdB, de manera de reducir totalmente los falsos positivos y evitar el tratamiento con DpnI. Otra mejora que consideramos llevar a cabo es la de adicionar a los vectores que contienen las fusiones con DsbC y MBP sus péptidos señal nativos para dirigir la fusión al periplasma mediante el mecanismo de translocación Sec, y permitir la formación de los S-S en este compartimento. De esta manera adicionaríamos 4 nuevos vectores a la serie ya construida.

Además de esto pensamos incorporar una segunda serie de 12 vectores, que presente niveles de expresión maximizados, mediante la optimización del RBS. Como se mencionó anteriormente, se ha visto que es posible diseñar y alterar las secuencias de unión al ribosoma (RBS) mediante un programa informático, para aumentar o disminuir las tasas de traducción (40, 41). Utilizando el programa disponible, se diseño un RBS sintético donde las tasas de traducción están maximizadas y que será introducido en la serie de vectores.

Finalmente, hemos visto que al aumentar las proteínas de fusión a evaluar, se aumentan también las probabilidades de lograr expresar en forma soluble la proteína blanco (78). Por lo tanto hemos considerado la posibilidad de incorporar a la serie de vectores una proteína de fusión adicional como ser la proteína periplásmica Spy (15.8 kDa) de *E. coli.* Esta mostró ser capaz de impedir la agregación y ayudar en la renaturalización proteica (304). En este sentido, ya clonamos la proteína para ser utilizada como proteína de fusión en la serie de vectores, extendiéndola ahora a 14 vectores.

En conclusión en esta primera parte del trabajo de Tesis se logró generar una serie de 12 vectores los cuales permiten un clonado en paralelo e independiente de la secuencia del gen de interés. Nuestros resultados muestran la capacidad de mejorar y facilitar en gran medida

esta etapa y proponen a una nueva proteína de fusión (CelDnc) como una herramienta de gran utilidad para incrementar la solubilidad/estabilidad de otras proteínas blanco. Estos resultados dieron origen a una reciente publicación en la revista Frontiers in Microbiology (5).

Una vez expresadas de forma soluble y homogénea, las PRs pueden ser utilizadas en una gran variedad de aplicaciones donde un área importante, es el de las proteínas de unión. Estas han sido utilizadas en aplicaciones biotecnológicas y también como reactivos para el estudio de procesos biológicos. Además, en la última década su uso en el campo clínico ha tenido una gran relevancia y crecimiento (6, 9). El tener por ejemplo, una proteína capaz de reconocer y unir específicamente un blanco terapéutico, puede servir tanto como herramienta para el diagnóstico (radioinmunodiagnosis *in vivo*, mediante acoplamiento con radionucleídos) así como también para el tratamiento de la enfermedad (conjugación con citotoxinas, o dominios para la interacción celular) (146, 148). Dentro de las proteínas de unión, las más ampliamente utilizadas corresponden a los AcMo, que han sido utilizados en el diagnóstico y tratamiento de numerosas patologías humanas (9, 146). Estas moléculas no sólo han mostrado ser efectivas en el reconocimiento de los blancos terapéuticos sino que forman parte de un mercado que genera miles de millones de dólares anuales (305). Sin embargo dependiendo de las aplicaciones, los AcMo pueden presentar algunas desventajas. Entre estas destacamos su gran tamaño que disminuye la penetración en tumores sólidos y larga vida media en suero, lo cual afecta el contraste en aplicaciones de imagenología (6). Además debido a la complejidad en su estructura química (presencia de S-S y sitios glicosilados), su producción a gran escala es compleja y costosa (6, 10). Versiones más pequeñas o simples de los AcMo, como los fragmentos Fab y scFv, han sido desarrolladas para sobrepasar alguna de las desventajas mencionadas. Si bien estos fragmentos se han producido con éxito en *E. coli*, su estabilidad aún depende de la correcta formación de puentes de azufre. Además algunos fragmentos de AcMo, tienden a agregar especialmente cuando están fusionados a dominios adicionales, por ejemplo para alcanzar una eficacia terapéutica o para su detección por lo que en estos casos su producción debe ser llevada a cabo en sistemas eucariotas (150).

Una de las razones de por qué los AcMo son las moléculas de unión más exitosas en la ciencia biomédica, es el hecho de que hasta hace 20 años el sistema inmune era la única fuente de diversidad molecular por la que la especificidad podía ser dirigida y seleccionada hacia un blanco específico. Actualmente, varias metodologías para la generación *in vitro* de grandes repertorios moleculares y selección de proteínas de alta afinidad y especificidad para un blanco determinado, han sido desarrolladas. El avance de estas metodologías de diversificación y selección derivó, en tan solo dos décadas, en la generación de más de 50

plegamientos proteicos diferentes o “scaffolds”, que sirven como proteínas de unión alternativas a los AcMo (152). Estas nuevas moléculas de unión o “binders”, combinan la fina especificidad y afinidad de los AcMo, con altos niveles de expresión en *E. coli*, falta de modificaciones post-traduccionales, estabilidad térmica e incluso química en amplios rangos de pH (150, 192). Es así que muchas de estas proteínas ya se encuentran en fases clínicas avanzadas contra patologías humanas como ser tratamiento contra cáncer, edema macular diabético, artritis reumatoide y osteoporosis entre otros (6). Esto demuestra el potencial de estos nuevos “scaffolds” para ser utilizados como una herramienta complementaria a la ya existente de los AcMo.

El éxito de una molécula de unión, radica en que sea capaz de unir de manera específica diferentes tipos de topologías como las que están presentes en las superficies proteicas. En este sentido, las moléculas pequeñas tienen la ventaja de poder acceder a regiones que pueden no ser muy accesibles para un AcMo como ser cavidades profundas (7, 8). Sin embargo, estas cavidades pueden ser muy conservadas y estar presentes en varias proteínas dentro del mismo organismo llevando a una inespecificidad de unión de las moléculas pequeñas ya que realizan pocos contactos con el blanco (267). Por lo tanto es necesario lograr un punto medio en donde se realicen un número de contactos suficientes como para conferir especificidad al interaccionar también con residuos no conservados, y a la vez una geometría tal que permita el acceso e interacción del sitio de unión con los residuos que se encuentran en la profundidad de las cavidades y que pueden ser los responsables por ejemplo de una actividad catalítica que se quiere inhibir. Además, dependiendo del blanco seleccionado, puede ser necesario en algunos casos unir superficies planas (de interacción proteína-proteína) y en otros cavidades profundas (sitios activos de algunas enzimas), requiriendo la generación de una librería con una diversidad estructural tal que pueda cumplir con ambos requerimientos.

En la segunda parte de este trabajo de Tesis nos abocamos a evaluar la capacidad de un scaffold de contener superficies de unión estructuralmente diferentes capaces de unir topologías diversas mediante distintos mecanismos de unión.

En estudios previos se demostró que las Afitinas derivadas de Sac7d de *S. acidocaldarius,* son capaces de unir diferentes epitopes del mismo blanco vía dos modos de interacción distintos Uno involucrando únicamente una superficie plana (L1, 14 residuos randomizados) y otro involucrando una superficie plana y 2 loops cortos (L2, 10 residuos randomizados) (192). Basados en estos resultados, se generó una nueva superficie de unión en Sac7d al adicionar 4

residuos randomizados entre en el loop que conecta las hebras β3 y β4 que se demostró podía participar en el reconocimiento de la IgG humana en una afitina previamente aislada (192). De esta manera se generaron 2 librerías L3 (13 residuos randomizados) y L4 (formado por 13 residuos randomizados y 2 posiciones mutadas a codones NHK) que contienen el loop extendido randomizado.

Como prueba de concepto decidimos utilizar una proteína blanco que sea fácil de producir y cristalizar, tenga un sitio activo profundo y que sea termoestable, para poder evaluar la interacción en un amplio rango de temperaturas. Es así que seleccionamos la endoglicosidasa CelD de *C. thermocellum* (EC 3.2.1.4), la cual cumple con las características mencionadas. Utilizando las diferentes librerías disponibles de Sac7d se lograron obtener proteínas de unión para CelD con afinidades determinadas por ITC del orden de nanomolar tanto a 25˚C como a 60˚C, únicamente con las librerías conteniendo el loop randomizado (L3 y L4). Utilizando 2 afitinas anti-CelD seleccionadas (E12 y H3) y una afitina (H4) capaz de unir a la glicosidasa lisozima (EC3.2.1.17) aislada previamente con la librería L1 (12), se encontró que eran inhibidores de sus respectivos blancos con una Ki en el orden de nanomolar. Además estos 3 inhibidores presentaron estabilidad térmica por DSC y no presentaron reacción cruzada por ITC y ELISA. Más aún, las afitinas anti-CelD eran capaces de inhibir esta enzima también a 60˚C a pesar de haber sido seleccionadas a 4˚C, por lo que las afitinas podrían ser una herramienta útil por ejemplo para investigaciones básicas *in vivo* de procesos biológicos con microorganismos termófilos. Las glicosidasas utilizadas en este estudio son capaces de hidrolizar enlaces O-glucosido y tienen sitios catalíticos profundos que utilizan 2 mecanismos enzimáticos diferentes. CelD cataliza la endohidrólisis de los enlaces (1→4) -β-D-glucosídicos presentes en la celulosa de los vegetales mediante un mecanismo de inversión de la configuración del C anomérico mientras que la lisozima cataliza la endohidrólisis de los enlaces (1→4)-β entre residuos de ácido *N*-acetilmurámico y *N*-acetyl-D-glucosamina del peptidoglicano bacteriano mediante un mecanismo de retención (306).

Se lograron resolver las estructuras cristalográficas de los diferentes complejos a alta resolución, revelando que el epitope reconocido en las dos enzimas está localizado en la cavidad catalítica. De esta manera se explica la inhibición en ambos casos debido al bloqueo del acceso al sitio activo. Más aún, las estructuras muestran una interacción directa por puentes de hidrógeno y puentes salinos con residuos catalíticamente importantes en ambas enzimas, bloqueando así su actividad. Mientras para el caso de la unión con la lisozima, la afitina se extiende a lo largo de la superficie cubriendo el sitio activo, en el caso de la unión con CelD, el loop extendido de las afitinas penetra en la cavidad del sitio activo para realizar las

interacciones. Las superficies de contacto de los diferentes complejos (desde 1317 $Å^2$ a 1749 $Å^2$) son comparables con las superficies encontradas en interacciones proteína-proteína (307). Por lo tanto, las estructuras muestran que es posible generar librerías capaces de unir blancos, mediante dos modos de interacción distintos, utilizando residuos randomizados en hojas β (afitina anti-lisozima) o una combinación de residuos presentes en hojas β y un loop extendido (afitinas anti-CelD). Estos últimos presentan un paratopo convexo capaz de penetrar en cavidades profundas. Por otro lado, las afitinas aisladas específicas por CelD, eran capaces de inhibir a la enzima a pesar de que en la selección no se aplicó ninguna presión selectiva para esto. Si bien más afitinas deberían ser aisladas para confirmarlo, no creemos que esto sea un caso fortuito, posiblemente la librería conteniendo el loop extendido tienda a unir superficies proteicas con fuerte curvatura. Este tipo de observaciones se confirmaron muy recientemente en un trabajo realizado por el grupo de Pluckthun. Utilizando un scaffold con un sitio de unión ya definido como ser las DARPins, estos autores generaron un nuevo repertorio con propiedades de unión diferentes al repertorio original (176). Este nuevo repertorio contenía un loop extendido randomizado en uno de las repeticiones de anquirina (LoopDARPins), formando un paratopo con estructura convexa, donde se vio (como en el caso de CelD), una tendencia a unir superficies cóncavas (176). Por otro lado, ya en trabajos no tan recientes, se había observado una tendencia en los nanobodies, que debido a la presencia de un loop CDR3 más aislado y extendido presentaban preferencias por unirse a cavidades profundas en las proteínas como la de los sitios activos de enzimas (10).

Otra característica interesante al analizar las estructuras obtenidas es que en la interacción afitina-glicosidasa, la afitina no interacciona únicamente con residuos catalíticos conservados, sino que se observan interacciones con residuos fuera del sitio activo que pueden contribuir con la especificidad de unión. En este sentido al analizar la secuencia de aminoácidos de 17 glicosidasas de la familia GH9 presentes en *C. thermocellum* (ATCC 27405), encontramos por ejemplo que en la interacción E12-CelD, la afitina interacciona con los residuos E353, Y354, V357, Y455, W538, P539, Y551 y E555. Tres de estos residuos no se encuentran en ninguna de las glicosidasas GH9 de *C. thermocellum* (E353, V357, W538), y P539 se encuentra solamente en 3 de ellas, pudiendo conferir especificidad de unión para CelD a la afitina E12.

Las estructuras obtenidas en este trabajo son las primeras estructuras de complejos affitina-proteína blanco, lo que nos permite determinar a nivel atómico cómo es que estas proteínas de unión logran interaccionar con sus blancos. Esta información nos permitirá además poder optimizar la librería al seleccionar nuevas posiciones para mutar y/o fijar posiciones que estaban mutadas pero no participan en la interacción.

BIBLIOGRAFÍA

1. Yang Z, Zhang L, Zhang Y, Zhang T, Feng Y, Lu X, Lan W, Wang J, Wu H, Cao C, Wang X. 2011. Highly efficient production of soluble proteins from insoluble inclusion bodies by a two-step-denaturing and refolding method. *PLoS One* 6: e22981
2. Correa A, Oppezzo P. 2011. Tuning different expression parameters to achieve soluble recombinant proteins in E. coli: advantages of high-throughput screening. *Biotechnol J* 6: 715-30
3. Correa A, Oppezzo P. 2014. *Overcoming the solubility problem in E. coli: available approaches for recombinant protein production (aceptado).* Methods in Molecular Biology-Humana Press
4. Unger T, Jacobovitch Y, Dantes A, Bernheim R, Peleg Y. 2010. Applications of the Restriction Free (RF) cloning procedure for molecular manipulations and protein expression. *J Struct Biol* 172: 34-44
5. Correa A, Ortega C, Obal G, Alzari P, Vincentelli R, Oppezzo PJ. 2014. Generation of a vector suite for protein solubility screening. *Frontiers in Microbiology* 5: 1-9
6. Wurch T, Pierre A, Depil S. 2012. Novel protein scaffolds as emerging therapeutic proteins: from discovery to clinical proof-of-concept. *Trends Biotechnol* 30: 575-82
7. van Montfort RL, Workman P. 2009. Structure-based design of molecular cancer therapeutics. *Trends Biotechnol* 27: 315-28
8. Verdine GL, Walensky LD. 2007. The challenge of drugging undruggable targets in cancer: lessons learned from targeting BCL-2 family members. *Clin Cancer Res* 13: 7264-70
9. An Z. 2010. Monoclonal antibodies - a proven and rapidly expanding therapeutic modality for human diseases. *Protein Cell* 1: 319-30
10. De Genst E, Silence K, Decanniere K, Conrath K, Loris R, Kinne J, Muyldermans S, Wyns L. 2006. Molecular basis for the preferential cleft recognition by dromedary heavy-chain antibodies. *Proc Natl Acad Sci U S A* 103: 4586-91
11. Mouratou B, Schaeffer F, Guilvout I, Tello-Manigne D, Pugsley AP, Alzari PM, Pecorari F. 2007. Remodeling a DNA-binding protein as a specific in vivo inhibitor of bacterial secretin PulD. *Proc Natl Acad Sci U S A* 104: 17983-8
12. Miranda FF, Brient-Litzler E, Zidane N, Pecorari F, Bedouelle H. 2011. Reagentless fluorescent biosensors from artificial families of antigen binding proteins. *Biosens Bioelectron* 26: 4184-90
13. Juy M, Amrt A, Alzari P, Poljak R, Claeyssens M. 1992. Three-dimensional structure of a thermostable bacterial cellulase. *Nature* 357
14. Delucas LJ, Hamrick D, Cosenza L, Nagy L, McCombs D, Bray T, Chait A, Stoops B, Belgovskiy A, William Wilson W, Parham M, Chernov N. 2005. Protein crystallization: virtual screening and optimization. *Prog Biophys Mol Biol* 88: 285-309
15. Chitarra V, Souchon H, Spinelli S, Juy M, Beguin P, Alzari PM. 1995. Multiple crystal forms of endoglucanase CelD: signal peptide residues modulate lattice formation. *J Mol Biol* 248: 225-32
16. FDA. 1982. Human insulin receives FDA approval. *FDA Drug Bull* 12: 18-9
17. Sorensen HP. 2010. Towards universal systems for recombinant gene expression. *Microb Cell Fact* 9: 27
18. Ferrer-Miralles N, Domingo-Espin J, Corchero JL, Vazquez E, Villaverde A. 2009. Microbial factories for recombinant pharmaceuticals. *Microb Cell Fact* 8: 17
19. Huang CJ, Lin H, Yang X. 2012. Industrial production of recombinant therapeutics in Escherichia coli and its recent advancements. *J Ind Microbiol Biotechnol* 39: 383-99

20. Brondyk WH. 2009. Selecting an appropriate method for expressing a recombinant protein. *Methods Enzymol* 463: 131-47
21. Widmann M, Christen P. 2000. Comparison of folding rates of homologous prokaryotic and eukaryotic proteins. *J Biol Chem* 275: 18619-22
22. Foit L, Morgan GJ, Kern MJ, Steimer LR, von Hacht AA, Titchmarsh J, Warriner SL, Radford SE, Bardwell JC. 2009. Optimizing protein stability in vivo. *Mol Cell* 36: 861-71
23. Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, de Castro E, Duvaud S, Flegel V, Fortier A, Gasteiger E, Grosdidier A, Hernandez C, Ioannidis V, Kuznetsov D, Liechti R, Moretti S, Mostaguir K, Redaschi N, Rossier G, Xenarios I, Stockinger H. 2012. ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res* 40: W597-603
24. Graslund S, Nordlund P, Weigelt J, Hallberg BM, Bray J, Gileadi O, Knapp S, Oppermann U, Arrowsmith C, Hui R, Ming J, dhe-Paganon S, Park HW, Savchenko A, Yee A, Edwards A, Vincentelli R, Cambillau C, Kim R, Kim SH, Rao Z, Shi Y, Terwilliger TC, Kim CY, Hung LW, Waldo GS, Peleg Y, Albeck S, Unger T, Dym O, Prilusky J, Sussman JL, Stevens RC, Lesley SA, Wilson IA, Joachimiak A, Collart F, Dementieva I, Donnelly MI, Eschenfeldt WH, Kim Y, Stols L, Wu R, Zhou M, Burley SK, Emtage JS, Sauder JM, Thompson D, Bain K, Luz J, Gheyi T, Zhang F, Atwell S, Almo SC, Bonanno JB, Fiser A, Swaminathan S, Studier FW, Chance MR, Sali A, Acton TB, Xiao R, Zhao L, Ma LC, Hunt JF, Tong L, Cunningham K, Inouye M, Anderson S, Janjua H, Shastry R, Ho CK, Wang D, Wang H, Jiang M, Montelione GT, Stuart DI, Owens RJ, Daenke S, Schutz A, Heinemann U, Yokoyama S, Bussow K, Gunsalus KC. 2008. Protein production and purification. *Nat Methods* 5: 135-46
25. Studier FW, Moffatt BA. 1986. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J Mol Biol* 189: 113-30
26. Dubendorff JW, Studier FW. 1991. Controlling basal expression in an inducible T7 expression system by blocking the target T7 promoter with lac repressor. *J Mol Biol* 219: 45-59
27. Studier FW, Rosenberg AH, Dunn JJ, Dubendorff JW. 1990. Use of T7 RNA polymerase to direct expression of cloned genes. *Methods Enzymol* 185: 60-89
28. Grossman TH, Kawasaki ES, Punreddy SR, Osburne MS. 1998. Spontaneous cAMP-dependent derepression of gene expression in stationary phase plays a role in recombinant expression instability. *Gene* 209: 95-103
29. Pan SH, Malcolm BA. 2000. Reduced background expression and improved plasmid stability with pET vectors in BL21 (DE3). *Biotechniques* 29: 1234-8
30. Brunner M, Bujard H. 1987. Promoter recognition and promoter strength in the Escherichia coli system. *Embo J* 6: 3139-44
31. Saida F, Uzan M, Odaert B, Bontems F. 2006. Expression of highly toxic genes in E. coli: special strategies and genetic tools. *Curr Protein Pept Sci* 7: 47-56
32. Guzman LM, Belin D, Carson MJ, Beckwith J. 1995. Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *J Bacteriol* 177: 4121-30
33. Goldstein J, Pollitt NS, Inouye M. 1990. Major cold shock protein of Escherichia coli. *Proc Natl Acad Sci U S A* 87: 283-7
34. Vasina JA, Baneyx F. 1996. Recombinant protein expression at low temperatures under the transcriptional control of the major Escherichia coli cold shock promoter cspA. *Appl Environ Microbiol* 62: 1444-7
35. Inouye S, Sahara Y. 2009. Expression and purification of the calcium binding photoprotein mitrocomin using ZZ-domain as a soluble partner in E. coli cells. *Protein Expr Purif* 66: 52-7
36. Villaverde A, Benito A, Viaplana E, Cubarsi R. 1993. Fine regulation of cI857-controlled gene expression in continuous culture of recombinant Escherichia coli by temperature. *Appl Environ Microbiol* 59: 3485-7

37. Valdez-Cruz NA, Caspeta L, Perez NO, Ramirez OT, Trujillo-Roldan MA. 2010. Production of recombinant proteins in E. coli by the heat inducible expression system based on the phage lambda pL and/or pR promoters. *Microb Cell Fact* 9: 18
38. Menart V, Jevsevar S, Vilar M, Trobis A, Pavko A. 2003. Constitutive versus thermoinducible expression of heterologous proteins in Escherichia coli based on strong PR,PL promoters from phage lambda. *Biotechnol Bioeng* 83: 181-90
39. Voges D, Watzele M, Nemetz C, Wizemann S, Buchberger B. 2004. Analyzing and enhancing mRNA translational efficiency in an Escherichia coli in vitro expression system. *Biochem Biophys Res Commun* 318: 601-14
40. Salis HM, Mirsky EA, Voigt CA. 2009. Automated design of synthetic ribosome binding sites to control protein expression. *Nat Biotechnol* 27: 946-50
41. Salis HM. 2011. The ribosome binding site calculator. *Methods Enzymol* 498: 19-42
42. Makino T, Skretas G, Georgiou G. 2011. Strain engineering for improved expression of recombinant proteins in bacteria. *Microb Cell Fact* 10: 32
43. Phillips TA, VanBogelen RA, Neidhardt FC. 1984. lon gene product of Escherichia coli is a heat-shock protein. *J Bacteriol* 159: 283-7
44. Grodberg J, Dunn JJ. 1988. ompT encodes the Escherichia coli outer membrane protease that cleaves T7 RNA polymerase during purification. *J Bacteriol* 170: 1245-53
45. Studier FW. 1991. Use of bacteriophage T7 lysozyme to improve an inducible T7 expression system. *J Mol Biol* 219: 37-44
46. Gustafsson C, Govindarajan S, Minshull J. 2004. Codon bias and heterologous protein expression. *Trends Biotechnol* 22: 346-53
47. Puigbo P, Guzman E, Romeu A, Garcia-Vallve S. 2007. OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. *Nucleic Acids Res* 35: W126-31
48. Villalobos A, Ness JE, Gustafsson C, Minshull J, Govindarajan S. 2006. Gene Designer: a synthetic biology tool for constructing artificial DNA segments. *BMC Bioinformatics* 7: 285
49. Maertens B, Spriestersbach A, von Groll U, Roth U, Kubicek J, Gerrits M, Graf M, Liss M, Daubert D, Wagner R, Schafer F. 2010. Gene optimization mechanisms: a multi-gene study reveals a high success rate of full-length human proteins expressed in Escherichia coli. *Protein Sci* 19: 1312-26
50. Burgess-Brown NA, Sharma S, Sobott F, Loenarz C, Oppermann U, Gileadi O. 2008. Codon optimization can improve expression of human genes in Escherichia coli: A multi-gene study. *Protein Expr Purif* 59: 94-102
51. Tegel H, Tourle S, Ottosson J, Persson A. 2010. Increased levels of recombinant human proteins with the Escherichia coli strain Rosetta(DE3). *Protein Expr Purif* 69: 159-67
52. Rosano GL, Ceccarelli EA. 2009. Rare codon content affects the solubility of recombinant proteins in a codon bias-adjusted Escherichia coli strain. *Microb Cell Fact* 8: 41
53. Marin M. 2008. Folding at the rhythm of the rare codon beat. *Biotechnol J* 3: 1047-57
54. Salinas G, Pellizza L, Margenat M, Flo M, Fernandez C. 2011. Tuned Escherichia coli as a host for the expression of disulfide-rich proteins. *Biotechnol J* 6: 686-99
55. Ferre F, Clote P. 2005. DiANNA: a web server for disulfide connectivity prediction. *Nucleic Acids Res* 33: W230-2
56. Pearce L, Morgan L, Lin TT, Hewamana S, Matthews RJ, Deaglio S, Rowntree C, Fegan C, Pepper C, Brennan P. 2010. Genetic modification of primary chronic lymphocytic leukemia cells with a lentivirus expressing CD38. *Haematologica* 95: 514-7
57. Berkmen M. 2012. Production of disulfide-bonded proteins in Escherichia coli. *Protein Expr Purif* 82: 240-51
58. de Marco A. 2009. Strategies for successful recombinant expression of disulfide bond-dependent proteins in Escherichia coli. *Microb Cell Fact* 8: 26

59. Schlegel S, Rujas E, Ytterberg AJ, Zubarev RA, Luirink J, de Gier JW. 2013. Optimizing heterologous protein production in the periplasm of E. coli by regulating gene expression levels. *Microb Cell Fact* 12: 24
60. Lee PA, Tullman-Ercek D, Georgiou G. 2006. The bacterial twin-arginine translocation pathway. *Annu Rev Microbiol* 60: 373-95
61. Natale P, Bruser T, Driessen AJ. 2008. Sec- and Tat-mediated protein secretion across the bacterial cytoplasmic membrane--distinct translocases and mechanisms. *Biochim Biophys Acta* 1778: 1735-56
62. Mergulhao FJ, Summers DK, Monteiro GA. 2005. Recombinant protein secretion in Escherichia coli. *Biotechnol Adv* 23: 177-202
63. Wagner S, Klepsch MM, Schlegel S, Appel A, Draheim R, Tarry M, Hogbom M, van Wijk KJ, Slotboom DJ, Persson JO, de Gier JW. 2008. Tuning Escherichia coli for membrane protein overexpression. *Proc Natl Acad Sci U S A* 105: 14371-6
64. Lobstein J, Emrich CA, Jeans C, Faulkner M, Riggs P, Berkmen M. 2012. SHuffle, a novel Escherichia coli protein expression strain capable of correctly folding disulfide bonded proteins in its cytoplasm. *Microb Cell Fact* 11: 56
65. Hatahet F, Nguyen VD, Salo KE, Ruddock LW. 2010. Disruption of reducing pathways is not essential for efficient disulfide bond formation in the cytoplasm of E. coli. *Microb Cell Fact* 9: 67
66. Nguyen VD, Hatahet F, Salo KE, Enlund E, Zhang C, Ruddock LW. 2011. Pre-expression of a sulfhydryl oxidase significantly increases the yields of eukaryotic disulfide bond containing proteins expressed in the cytoplasm of E.coli. *Microb Cell Fact* 10: 1
67. Nozach H, Fruchart-Gaillard C, Fenaille F, Beau F, Ramos OH, Douzi B, Saez NJ, Moutiez M, Servent D, Gondry M, Thai R, Cuniasse P, Vincentelli R, Dive V. 2013. High throughput screening identifies disulfide isomerase DsbC as a very efficient partner for recombinant expression of small disulfide-rich proteins in E. coli. *Microb Cell Fact* 12: 37
68. Ahram M, Litou ZI, Fang R, Al-Tawallbeh G. 2006. Estimation of membrane proteins in the human proteome. *In Silico Biol* 6: 379-86
69. Scott DJ, Kummer L, Tremmel D, Pluckthun A. 2013. Stabilizing membrane proteins through protein engineering. *Curr Opin Chem Biol* 17: 427-35
70. Freigassner M, Pichler H, Glieder A. 2009. Tuning microbial hosts for membrane protein production. *Microb Cell Fact* 8: 69
71. Kawate T, Gouaux E. 2006. Fluorescence-detection size-exclusion chromatography for precrystallization screening of integral membrane proteins. *Structure* 14: 673-81
72. Backmark AE, Olivier N, Snijder A, Gordon E, Dekker N, Ferguson AD. 2013. Fluorescent probe for high-throughput screening of membrane protein expression. *Protein Sci* 22: 1124-32
73. Miroux B, Walker JE. 1996. Over-production of proteins in Escherichia coli: mutant hosts that allow synthesis of some membrane proteins and globular proteins at high levels. *J Mol Biol* 260: 289-98
74. Tate CG, Haase J, Baker C, Boorsma M, Magnani F, Vallis Y, Williams DC. 2003. Comparison of seven different heterologous protein expression systems for the production of the serotonin transporter. *Biochim Biophys Acta* 1610: 141-53
75. Yamanaka K. 1999. Cold shock response in Escherichia coli. *J Mol Microbiol Biotechnol* 1: 193-202
76. Vera A, Gonzalez-Montalban N, Aris A, Villaverde A. 2007. The conformational quality of insoluble recombinant proteins is enhanced at low growth temperatures. *Biotechnol Bioeng* 96: 1101-6
77. Studier FW. 2005. Protein production by auto-induction in high density shaking cultures. *Protein Expr Purif* 41: 207-34

78. Vincentelli R, Cimino A, Geerlof A, Kubo A, Satou Y, Cambillau C. 2011. High-throughput protein expression screening and purification in Escherichia coli. *Methods* 55: 65-72
79. Vincentelli R, Romier C. 2013. Expression in Escherichia coli: becoming faster and more complex. *Curr Opin Struct Biol* 23: 326-34
80. Hailu TT, Foit L, Bardwell JC. 2013. In vivo detection and quantification of chemicals that enhance protein stability. *Anal Biochem* 434: 181-6
81. Murphy MB, Doyle SA. 2005. High-throughput purification of hexahistidine-tagged proteins expressed in E. coli. *Methods Mol Biol* 310: 123-30
82. Zhu XQ, Li SX, He HJ, Yuan QS. 2005. On-column refolding of an insoluble His6-tagged recombinant EC-SOD overexpressed in Escherichia coli. *Acta Biochim Biophys Sin (Shanghai)* 37: 265-9
83. Li M, Su ZG, Janson JC. 2004. In vitro protein refolding by chromatographic procedures. *Protein Expr Purif* 33: 1-10
84. Magnusdottir A, Johansson I, Dahlgren LG, Nordlund P, Berglund H. 2009. Enabling IMAC purification of low abundance recombinant proteins from E. coli lysates. *Nat Methods* 6: 477-8
85. Bolanos-Garcia VM, Davies OR. 2006. Structural analysis and classification of native proteins from E. coli commonly co-purified by immobilised metal affinity chromatography. *Biochim Biophys Acta* 1760: 1304-13
86. Robichon C, Luo J, Causey TB, Benner JS, Samuelson JC. 2011. Engineering Escherichia coli BL21(DE3) derivative strains to minimize E. coli protein contamination after purification by immobilized metal affinity chromatography. *Appl Environ Microbiol* 77: 4634-46
87. Andersen KR, Leksa NC, Schwartz TU. 2013. Optimized E. coli expression strain LOBSTR eliminates common contaminants from His-tag purification. *Proteins*
88. Schmidt TG, Skerra A. 2007. The Strep-tag system for one-step purification and high-affinity detection or capturing of proteins. *Nat Protoc* 2: 1528-35
89. Lichty JJ, Malecki JL, Agnew HD, Michelson-Horowitz DJ, Tan S. 2005. Comparison of affinity tags for protein purification. *Protein Expr Purif* 41: 98-105
90. Hammarstrom M, Hellgren N, van Den Berg S, Berglund H, Hard T. 2002. Rapid screening for improved solubility of small human proteins produced as fusion proteins in Escherichia coli. *Protein Sci* 11: 313-21
91. Esposito D, Chatterjee DK. 2006. Enhancement of soluble protein expression through the use of fusion tags. *Curr Opin Biotechnol* 17: 353-8
92. Young CL, Britton ZT, Robinson AS. 2012. Recombinant protein expression and purification: a comprehensive review of affinity tags and microbial applications. *Biotechnol J* 7: 620-34
93. Pattenden LK, Thomas WG. 2008. Amylose affinity chromatography of maltose-binding protein: purification by both native and novel matrix-assisted dialysis refolding methods. *Methods Mol Biol* 421: 169-89
94. Dyson MR, Shadbolt SP, Vincent KJ, Perera RL, McCafferty J. 2004. Production of soluble mammalian proteins in Escherichia coli: identification of protein features that correlate with successful expression. *BMC Biotechnol* 4: 32
95. Kapust RB, Waugh DS. 1999. Escherichia coli maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci* 8: 1668-74
96. Cho HJ, Lee Y, Chang RS, Hahm MS, Kim MK, Kim YB, Oh YK. 2008. Maltose binding protein facilitates high-level expression and functional purification of the chemokines RANTES and SDF-1alpha from Escherichia coli. *Protein Expr Purif* 60: 37-45
97. Smith DB, Johnson KS. 1988. Single-step purification of polypeptides expressed in Escherichia coli as fusions with glutathione S-transferase. *Gene* 67: 31-40

98. Lunn CA, Kathju S, Wallace BJ, Kushner SR, Pigiet V. 1984. Amplification and purification of plasmid-encoded thioredoxin from Escherichia coli K12. *J Biol Chem* 259: 10469-74
99. LaVallie ER, DiBlasio EA, Kovacic S, Grant KL, Schendel PF, McCoy JM. 1993. A thioredoxin gene fusion expression system that circumvents inclusion body formation in the E. coli cytoplasm. *Biotechnology (N Y)* 11: 187-93
100. LaVallie ER, Lu Z, Diblasio-Smith EA, Collins-Racie LA, McCoy JM. 2000. Thioredoxin as a fusion partner for production of soluble recombinant proteins in Escherichia coli. *Methods Enzymol* 326: 322-40
101. Kim S, Lee SB. 2008. Soluble expression of archaeal proteins in Escherichia coli by using fusion-partners. *Protein Expr Purif* 62: 116-9
102. Zhang Z, Li ZH, Wang F, Fang M, Yin CC, Zhou ZY, Lin Q, Huang HL. 2002. Overexpression of DsbC and DsbG markedly improves soluble and functional expression of single-chain Fv antibodies in Escherichia coli. *Protein Expr Purif* 26: 218-28
103. Marblestone JG, Edavettal SC, Lim Y, Lim P, Zuo X, Butt TR. 2006. Comparison of SUMO fusion technology with traditional gene fusion systems: enhanced expression and solubility with SUMO. *Protein Sci* 15: 182-9
104. Malakhov MP, Mattern MR, Malakhova OA, Drinker M, Weeks SD, Butt TR. 2004. SUMO fusions and SUMO-specific protease for efficient expression and purification of proteins. *J Struct Funct Genomics* 5: 75-86
105. Butt TR, Edavettal SC, Hall JP, Mattern MR. 2005. SUMO fusion technology for difficult-to-express proteins. *Protein Expr Purif* 43: 1-9
106. De Marco V, Stier G, Blandin S, de Marco A. 2004. The solubility and stability of recombinant proteins are increased by their fusion to NusA. *Biochem Biophys Res Commun* 322: 766-71
107. Nallamsetty S, Waugh DS. 2006. Solubility-enhancing proteins MBP and NusA play a passive role in the folding of their fusion partners. *Protein Expr Purif* 45: 175-82
108. Moon AF, Mueller GA, Zhong X, Pedersen LC. 2010. A synergistic approach to protein crystallization: combination of a fixed-arm carrier with surface entropy reduction. *Protein Sci* 19: 901-13
109. Suzuki N, Hiraki M, Yamada Y, Matsugaki N, Igarashi N, Kato R, Dikic I, Drew D, Iwata S, Wakatsuki S, Kawasaki M. 2010. Crystallization of small proteins assisted by green fluorescent protein. *Acta Crystallogr D Biol Crystallogr* 66: 1059-66
110. Smyth DR, Mrozkiewicz MK, McGrath WJ, Listwan P, Kobe B. 2003. Crystal structures of fusion proteins with large-affinity tags. *Protein Sci* 12: 1313-22
111. Corsini L, Hothorn M, Scheffzek K, Sattler M, Stier G. 2008. Thioredoxin as a fusion tag for carrier-driven crystallization. *Protein Sci* 17: 2070-9
112. van den Berg S, Lofdahl PA, Hard T, Berglund H. 2006. Improved solubility of TEV protease by directed evolution. *J Biotechnol* 121: 291-8
113. Klint JK, Senff S, Saez NJ, Seshadri R, Lau HY, Bende NS, Undheim EA, Rash LD, Mobli M, King GF. 2013. Production of recombinant disulfide-rich venom peptides for structural and functional analysis via expression in the periplasm of E. coli. *PLoS One* 8: e63865
114. Acton TB, Gunsalus KC, Xiao R, Ma LC, Aramini J, Baran MC, Chiang YW, Climent T, Cooper B, Denissova NG, Douglas SM, Everett JK, Ho CK, Macapagal D, Rajan PK, Shastry R, Shih LY, Swapna GV, Wilson M, Wu M, Gerstein M, Inouye M, Hunt JF, Montelione GT. 2005. Robotic cloning and Protein Production Platform of the Northeast Structural Genomics Consortium. *Methods Enzymol* 394: 210-43
115. Xiao R, Anderson S, Aramini J, Belote R, Buchwald WA, Ciccosanti C, Conover K, Everett JK, Hamilton K, Huang YJ, Janjua H, Jiang M, Kornhaber GJ, Lee DY, Locke JY, Ma LC, Maglaqui M, Mao L, Mitra S, Patel D, Rossi P, Sahdev S, Sharma S, Shastry R, Swapna GV, Tong SN, Wang D, Wang H, Zhao L, Montelione GT, Acton TB. 2010. The high-

throughput protein sample production platform of the Northeast Structural Genomics Consortium. *J Struct Biol* 172: 21-33
116. Vincentelli R, Canaan S, Campanacci V, Valencia C, Maurin D, Frassinetti F, Scappucini-Calvo L, Bourne Y, Cambillau C, Bignon C. 2004. High-throughput automated refolding screening of inclusion bodies. *Protein Sci* 13: 2782-92
117. Eshaghi S, Hedren M, Nasser MI, Hammarberg T, Thornell A, Nordlund P. 2005. An efficient strategy for high-throughput expression screening of recombinant integral membrane proteins. *Protein Sci* 14: 676-83
118. Koehn J, Hunt I. 2009. High-Throughput Protein Production (HTPP): a review of enabling technologies to expedite protein production. *Methods Mol Biol* 498: 1-18
119. Lin CT, Moore PA, Kery V. 2009. Automated 96-well purification of hexahistidine-tagged recombinant proteins on MagneHis Ni(2)+-particles. *Methods Mol Biol* 498: 129-41
120. Schafer F, Romer U, Emmerlich M, Blumer J, Lubenow H, Steinert K. 2002. Automated high-throughput purification of 6xHis-tagged proteins. *J Biomol Tech* 13: 131-42
121. Ventura S, Villaverde A. 2006. Protein quality in bacterial inclusion bodies. *Trends Biotechnol* 24: 179-85
122. Dechavanne V, Barrillat N, Borlat F, Hermant A, Magnenat L, Paquet M, Antonsson B, Chevalet L. 2010. A high-throughput protein refolding screen in 96-well format combined with design of experiments to optimize the refolding conditions. *Protein Expr Purif* 75: 192-203
123. Clark EDB. 1998. Refolding of recombinant proteins. *Curr Opin Biotechnol* 9: 157-63
124. Achmuller C, Kaar W, Ahrer K, Wechner P, Hahn R, Werther F, Schmidinger H, Cserjan-Puschmann M, Clementschitsch F, Striedner G, Bayer K, Jungbauer A, Auer B. 2007. N(pro) fusion technology to produce proteins with authentic N termini in E. coli. *Nat Methods* 4: 1037-43
125. Ke T, Liang S, Huang J, Mao H, Chen J, Dong C, Liu S, Kang J, Liu D, Ma X. 2012. A novel PCR-based method for high throughput prokaryotic expression of antimicrobial peptide genes. *BMC Biotechnol* 12: 10
126. Tokatlidis K, Dhurjati P, Millet J, Beguin P, Aubert JP. 1991. High activity of inclusion bodies formed in Escherichia coli overproducing Clostridium thermocellum endoglucanase D. *FEBS Lett* 282: 205-8
127. Garcia-Fruitos E, Gonzalez-Montalban N, Morell M, Vera A, Ferraz RM, Aris A, Ventura S, Villaverde A. 2005. Aggregation as bacterial inclusion bodies does not imply inactivation of enzymes and fluorescent proteins. *Microb Cell Fact* 4: 27
128. de Groot NS, Ventura S. 2006. Protein activity in bacterial inclusion bodies correlates with predicted aggregation rates. *J Biotechnol* 125: 110-3
129. Peternel S, Grdadolnik J, Gaberc-Porekar V, Komel R. 2008. Engineering inclusion bodies for non denaturing extraction of functional proteins. *Microb Cell Fact* 7: 34
130. Garcia-Fruitos E. 2010. Inclusion bodies: a new concept. *Microb Cell Fact* 9: 80
131. Garcia-Fruitos E, Vazquez E, Diez-Gil C, Corchero JL, Seras-Franzoso J, Ratera I, Veciana J, Villaverde A. 2012. Bacterial inclusion bodies: making gold from waste. *Trends Biotechnol* 30: 65-70
132. Villaverde A, Garcia-Fruitos E, Rinas U, Seras-Franzoso J, Kosoy A, Corchero JL, Vazquez E. 2012. Packaging protein drugs as bacterial inclusion bodies for therapeutic applications. *Microb Cell Fact* 11: 76
133. Eijsink VG, Bjork A, Gaseidnes S, Sirevag R, Synstad B, van den Burg B, Vriend G. 2004. Rational engineering of enzyme stability. *J Biotechnol* 113: 105-20
134. Stemmer WP. 1994. Rapid evolution of a protein in vitro by DNA shuffling. *Nature* 370: 389-91
135. Roodveldt C, Aharoni A, Tawfik DS. 2005. Directed evolution of proteins for heterologous expression and stability. *Curr Opin Struct Biol* 15: 50-6

136. Waldo GS, Standish BM, Berendzen J, Terwilliger TC. 1999. Rapid protein-folding assay using green fluorescent protein. *Nat Biotechnol* 17: 691-5
137. Pedelacq JD, Piltch E, Liong EC, Berendzen J, Kim CY, Rho BS, Park MS, Terwilliger TC, Waldo GS. 2002. Engineering soluble proteins for structural genomics. *Nat Biotechnol* 20: 927-32
138. Cabantous S, Terwilliger TC, Waldo GS. 2005. Protein tagging and detection with engineered self-assembling fragments of green fluorescent protein. *Nat Biotechnol* 23: 102-7
139. Pedelacq JD, Cabantous S, Tran T, Terwilliger TC, Waldo GS. 2006. Engineering and characterization of a superfolder green fluorescent protein. *Nat Biotechnol* 24: 79-88
140. Maxwell KL, Mittermaier AK, Forman-Kay JD, Davidson AR. 1999. A simple in vivo assay for increased protein solubility. *Protein Sci* 8: 1908-11
141. Sieber V, Martinez CA, Arnold FH. 2001. Libraries of hybrid proteins from distantly related sequences. *Nat Biotechnol* 19: 456-60
142. Dahlroth SL, Nordlund P, Cornvik T. 2006. Colony filtration blotting for screening soluble expression in Escherichia coli. *Nat Protoc* 1: 253-8
143. Cornvik T, Dahlroth SL, Magnusdottir A, Herman MD, Knaust R, Ekberg M, Nordlund P. 2005. Colony filtration blot: a new screening method for soluble protein expression in Escherichia coli. *Nat Methods* 2: 507-9
144. Low C, Moberg P, Quistgaard EM, Hedren M, Guettou F, Frauenfeld J, Haneskog L, Nordlund P. 2013. High-throughput analytical gel filtration screening of integral membrane proteins for structural studies. *Biochim Biophys Acta* 1830: 3497-508
145. Sala E, de Marco A. 2010. Screening optimized protein purification protocols by coupling small-scale expression and mini-size exclusion chromatography. *Protein Expr Purif* 74: 231-5
146. Iyer U, Kadambi VJ. 2011. Antibody drug conjugates - Trojan horses in the war on cancer. *J Pharmacol Toxicol Methods* 64: 207-12
147. Farid SS. 2007. Process economics of industrial monoclonal antibody manufacture. *J Chromatogr B Analyt Technol Biomed Life Sci* 848: 8-18
148. Eggenstein E, Eichinger A, Kim HJ, Skerra A. 2014. Structure-guided engineering of Anticalins with improved binding behavior and biochemical characteristics for application in radio-immuno imaging and/or therapy. *J Struct Biol* 185: 203-14
149. Pancer Z, Amemiya CT, Ehrhardt GR, Ceitlin J, Gartland GL, Cooper MD. 2004. Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. *Nature* 430: 174-80
150. Binz HK, Amstutz P, Pluckthun A. 2005. Engineering novel binding proteins from nonimmunoglobulin domains. *Nat Biotechnol* 23: 1257-68
151. Binz HK, Pluckthun A. 2005. Engineered proteins as specific binding reagents. *Curr Opin Biotechnol* 16: 459-69
152. Gebauer M, Skerra A. 2009. Engineered protein scaffolds as next-generation antibody therapeutics. *Curr Opin Chem Biol* 13: 245-55
153. Weidle UH, Auer J, Brinkmann U, Georges G, Tiefenthaler G. 2013. The emerging role of new protein scaffold-based agents for treatment of cancer. *Cancer Genomics Proteomics* 10: 155-68
154. Virnekas B, Ge L, Pluckthun A, Schneider KC, Wellnhofer G, Moroney SE. 1994. Trinucleotide phosphoramidites: ideal reagents for the synthesis of mixed oligonucleotides for random mutagenesis. *Nucleic Acids Res* 22: 5600-7
155. Gronwall C, Stahl S. 2009. Engineered affinity proteins--generation and applications. *J Biotechnol* 140: 254-69
156. Forrer P, Stumpp MT, Binz HK, Pluckthun A. 2003. A novel strategy to design binding molecules harnessing the modular nature of repeat proteins. *FEBS Lett* 539: 2-6

157. Forrer P, Binz HK, Stumpp MT, Pluckthun A. 2004. Consensus design of repeat proteins. *Chembiochem* 5: 183-9
158. Lofblom J, Feldwisch J, Tolmachev V, Carlsson J, Stahl S, Frejd FY. 2010. Affibody molecules: engineered proteins for therapeutic, diagnostic and biotechnological applications. *FEBS Lett* 584: 2670-80
159. Kohl A, Binz HK, Forrer P, Stumpp MT, Pluckthun A, Grutter MG. 2003. Designed to be stable: crystal structure of a consensus ankyrin repeat protein. *Proc Natl Acad Sci U S A* 100: 1700-5
160. Gebauer M, Skerra A. 2012. Anticalins small engineered binding proteins based on the lipocalin scaffold. *Methods Enzymol* 503: 157-88
161. Dennis MS, Herzka A, Lazarus RA. 1995. Potent and selective Kunitz domain inhibitors of plasma kallikrein designed by phage display. *J Biol Chem* 270: 25411-7
162. Muyldermans S. 2013. Nanobodies: natural single-domain antibodies. *Annu Rev Biochem* 82: 775-97
163. Nord K, Gunneriusson E, Ringdahl J, Stahl S, Uhlen M, Nygren PA. 1997. Binding proteins selected from combinatorial libraries of an alpha-helical bacterial receptor domain. *Nat Biotechnol* 15: 772-7
164. Nygren PA. 2008. Alternative binding proteins: affibody binding proteins developed from a small three-helix bundle scaffold. *FEBS J* 275: 2668-76
165. Tolmachev V, Rosik D, Wallberg H, Sjoberg A, Sandstrom M, Hansson M, Wennborg A, Orlova A. 2010. Imaging of EGFR expression in murine xenografts using site-specifically labelled anti-EGFR 111In-DOTA-Z EGFR:2377 Affibody molecule: aspect of the injected tracer amount. *Eur J Nucl Med Mol Imaging* 37: 613-22
166. Engfeldt T, Renberg B, Brumer H, Nygren PA, Karlstrom AE. 2005. Chemical synthesis of triple-labelled three-helix bundle binding proteins for specific fluorescent detection of unlabelled protein. *Chembiochem* 6: 1043-50
167. Orlova A, Tolmachev V, Pehrson R, Lindborg M, Tran T, Sandstrom M, Nilsson FY, Wennborg A, Abrahmsen L, Feldwisch J. 2007. Synthetic affibody molecules: a novel class of affinity ligands for molecular imaging of HER2-expressing malignant tumors. *Cancer Res* 67: 2178-86
168. Baum RP, Prasad V, Muller D, Schuchardt C, Orlova A, Wennborg A, Tolmachev V, Feldwisch J. 2010. Molecular imaging of HER2-expressing malignant tumors in breast cancer patients using synthetic 111In- or 68Ga-labeled affibody molecules. *J Nucl Med* 51: 892-7
169. Nord K, Gunneriusson E, Uhlen M, Nygren PA. 2000. Ligands selected from combinatorial libraries of protein A for use in affinity capture of apolipoprotein A-1M and taq DNA polymerase. *J Biotechnol* 80: 45-54
170. Nord K, Nord O, Uhlen M, Kelley B, Ljungqvist C, Nygren PA. 2001. Recombinant human factor VIII-specific affinity ligands selected from phage-displayed combinatorial libraries of protein A. *Eur J Biochem* 268: 4269-77
171. Ronnmark J, Gronlund H, Uhlen M, Nygren PA. 2002. Human immunoglobulin A (IgA)-specific ligands from combinatorial engineering of protein A. *Eur J Biochem* 269: 2647-55
172. Ronnmark J, Kampf C, Asplund A, Hoiden-Guthenberg I, Wester K, Ponten F, Uhlen M, Nygren PA. 2003. Affibody-beta-galactosidase immunoconjugates produced as soluble fusion proteins in the Escherichia coli cytosol. *J Immunol Methods* 281: 149-60
173. Lundberg E, Hoiden-Guthenberg I, Larsson B, Uhlen M, Graslund T. 2007. Site-specifically conjugated anti-HER2 Affibody molecules as one-step reagents for target expression analyses on cells and xenograft samples. *J Immunol Methods* 319: 53-63
174. Binz HK, Stumpp MT, Forrer P, Amstutz P, Pluckthun A. 2003. Designing repeat proteins: well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. *J Mol Biol* 332: 489-503

175. Zahnd C, Wyler E, Schwenk JM, Steiner D, Lawrence MC, McKern NM, Pecorari F, Ward CW, Joos TO, Pluckthun A. 2007. A designed ankyrin repeat protein evolved to picomolar affinity to Her2. *J Mol Biol* 369: 1015-28
176. Schilling J, Schoppe J, Pluckthun A. 2014. From DARPins to LoopDARPins: Novel LoopDARPin design allows the selection of low picomolar binders in a single round of ribosome display. *J Mol Biol*
177. Campochiaro PA, Channa R, Berger BB, Heier JS, Brown DM, Fiedler U, Hepp J, Stumpp MT. 2013. Treatment of diabetic macular edema with a designed ankyrin repeat protein that binds vascular endothelial growth factor: a phase I/II study. *Am J Ophthalmol* 155: 697-704, e1-2
178. Dreier B, Honegger A, Hess C, Nagy-Davidescu G, Mittl PR, Grutter MG, Belousova N, Mikheeva G, Krasnykh V, Pluckthun A. 2013. Development of a generic adenovirus delivery system based on structure-guided design of bispecific trimeric DARPin adapters. *Proc Natl Acad Sci U S A* 110: E869-77
179. Skerra A. 2008. Alternative binding proteins: anticalins - harnessing the structural plasticity of the lipocalin ligand pocket to engineer novel binding activities. *FEBS J* 275: 2677-83
180. Mross K, Richly H, Fischer R, Scharr D, Buchert M, Stern A, Gille H, Audoly LP, Scheulen ME. 2013. First-in-Human Phase I Study of PRS-050 (Angiocal), an Anticalin Targeting and Antagonizing VEGF-A, in Patients with Advanced Solid Tumors. *PLoS One* 8: e83232
181. Kim HJ, Eichinger A, Skerra A. 2009. High-affinity recognition of lanthanide(III) chelate complexes by a reprogrammed human lipocalin 2. *J Am Chem Soc* 131: 3565-76
182. Martello JL, Woytowish MR, Chambers H. 2012. Ecallantide for treatment of acute attacks of hereditary angioedema. *Am J Health Syst Pharm* 69: 651-7
183. van der Linden RH, Frenken LG, de Geus B, Harmsen MM, Ruuls RC, Stok W, de Ron L, Wilson S, Davis P, Verrips CT. 1999. Comparison of physical chemical properties of llama VHH antibody fragments and mouse monoclonal antibodies. *Biochim Biophys Acta* 1431: 37-46
184. Holliger P, Hudson PJ. 2005. Engineered antibody fragments and the rise of single domains. *Nat Biotechnol* 23: 1126-36
185. Rothbauer U, Zolghadr K, Tillib S, Nowak D, Schermelleh L, Gahl A, Backmann N, Conrath K, Muyldermans S, Cardoso MC, Leonhardt H. 2006. Targeting and tracing antigens in live cells with fluorescent nanobodies. *Nat Methods* 3: 887-9
186. Olichon A, Surrey T. 2007. Selection of genetically encoded fluorescent single domain antibodies engineered for efficient expression in Escherichia coli. *J Biol Chem* 282: 36314-20
187. Van de Broek B, Devoogdt N, D'Hollander A, Gijs HL, Jans K, Lagae L, Muyldermans S, Maes G, Borghs G. 2011. Specific cell targeting with nanobody conjugated branched gold nanoparticles for photothermal therapy. *ACS Nano* 5: 4319-28
188. Hmila I, Saerens D, Ben Abderrazek R, Vincke C, Abidi N, Benlasfar Z, Govaert J, El Ayeb M, Bouhaouala-Zahar B, Muyldermans S. 2010. A bispecific nanobody to provide full protection against lethal scorpion envenoming. *FASEB J* 24: 3479-89
189. Williams SC. 2014. Small nanobody drugs win big backing from pharma. *Nat Med* 19: 1355-6
190. Edmondson SP, Shriver JW. 2001. DNA binding proteins Sac7d and Sso7d from Sulfolobus. *Methods Enzymol* 334: 129-45
191. Buddelmeijer N, Krehenbrink M, Pecorari F, Pugsley AP. 2009. Type II secretion system secretin PulD localizes in clusters in the Escherichia coli outer membrane. *J Bacteriol* 191: 161-8
192. Behar G, Bellinzoni M, Maillasson M, Paillard-Laurance L, Alzari PM, He X, Mouratou B, Pecorari F. 2013. Tolerance of the archaeal Sac7d scaffold protein to alternative library

designs: characterization of anti-immunoglobulin G Affitins. *Protein Eng Des Sel* 26: 267-75
193. Smith GP. 1985. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* 228: 1315-7
194. Webster R. 2001. *Filamentous phage biology*: Cold Spring Harbor, NY
195. Boder ET, Wittrup KD. 1997. Yeast surface display for screening combinatorial polypeptide libraries. *Nat Biotechnol* 15: 553-7
196. Daugherty PS. 2007. Protein engineering with bacterial display. *Curr Opin Struct Biol* 17: 474-80
197. Pluckthun A. 2012. Ribosome display: a perspective. *Methods Mol Biol* 805: 3-28
198. Leemhuis H, Stein V, Griffiths AD, Hollfelder F. 2005. New genotype-phenotype linkages for directed evolution of functional proteins. *Curr Opin Struct Biol* 15: 472-8
199. Hanes J, Pluckthun A. 1997. In vitro selection and evolution of functional proteins by using ribosome display. *Proc Natl Acad Sci U S A* 94: 4937-42
200. Takahashi TT, Roberts RW. 2009. In vitro selection of protein and peptide libraries using mRNA display. *Methods Mol Biol* 535: 293-314
201. Yamaguchi J, Naimuddin M, Biyani M, Sasaki T, Machida M, Kubo T, Funatsu T, Husimi Y, Nemoto N. 2009. cDNA display: a novel screening method for functional disulfide-rich peptides by solid-phase synthesis and stabilization of mRNA-protein fusions. *Nucleic Acids Res* 37: e108
202. Hoess RH. 2001. Protein design and phage display. *Chem Rev* 101: 3205-18
203. Loset GA, Sandlie I. 2012. Next generation phage display by use of pVII and pIX as display scaffolds. *Methods* 58: 40-6
204. Mattheakis LC, Bhatt RR, Dower WJ. 1994. An in vitro polysome display system for identifying ligands from very large peptide libraries. *Proc Natl Acad Sci U S A* 91: 9022-6
205. Lipovsek D, Pluckthun A. 2004. In-vitro protein evolution by ribosome display and mRNA display. *J Immunol Methods* 290: 51-67
206. Dreier B, Pluckthun A. 2011. Ribosome display: a technology for selecting and evolving proteins from large libraries. *Methods Mol Biol* 687: 283-306
207. Zahnd C, Amstutz P, Pluckthun A. 2007. Ribosome display: selecting and evolving proteins in vitro that specifically bind to a target. *Nat Methods* 4: 269-79
208. Zahnd C, Sarkar CA, Pluckthun A. 2010. Computational analysis of off-rate selection experiments to optimize affinity maturation by directed evolution. *Protein Eng Des Sel* 23: 175-84
209. Yanagida H, Matsuura T, Yomo T. 2010. Ribosome display for rapid protein evolution by consecutive rounds of mutation and selection. *Methods Mol Biol* 634: 257-67
210. Ueda T, Kanamori T, Ohashi H. 2010. Ribosome display with the PURE technology. *Methods Mol Biol* 607: 219-25
211. Ohashi H, Shimizu Y, Ying BW, Ueda T. 2007. Efficient protein selection based on ribosome display system with purified components. *Biochem Biophys Res Commun* 352: 270-6
212. Matsuura T, Yanagida H, Ushioda J, Urabe I, Yomo T. 2007. Nascent chain, mRNA, and ribosome complexes generated by a pure translation system. *Biochem Biophys Res Commun* 352: 372-7
213. Evans MS, Ugrinov KG, Frese MA, Clark PL. 2005. Homogeneous stalled ribosome nascent chain complexes produced in vivo or in vitro. *Nat Methods* 2: 757-62
214. Wada A, Ito Y. 2009. The highly stabilized ribosome display selection of metal binding peptide aptamers. *Nucleic Acids Symp Ser (Oxf)*: 263-4
215. Ueno S, Kimura S, Ichiki T, Nemoto N. 2012. Improvement of a puromycin-linker to extend the selection target varieties in cDNA display method. *J Biotechnol* 162: 299-302

216. Gilbreth RN, Koide S. 2012. Structural insights for engineering binding proteins based on non-antibody scaffolds. *Curr Opin Struct Biol* 22: 413-20
217. Finkelstein AV, Janin J. 1989. The price of lost freedom: entropy of bimolecular complex formation. *Protein Eng* 3: 1-3
218. Tidor B, Karplus M. 1994. The contribution of vibrational entropy to molecular association. The dimerization of insulin. *J Mol Biol* 238: 405-14
219. Vekilov PG, Feeling-Taylor AR, Yau ST, Petsev D. 2002. Solvent entropy contribution to the free energy of protein crystallization. *Acta Crystallogr D Biol Crystallogr* 58: 1611-6
220. Bukowska MA, Grutter MG. 2013. New concepts and aids to facilitate crystallization. *Curr Opin Struct Biol* 23: 409-16
221. Sennhauser G, Grutter MG. 2008. Chaperone-assisted crystallography with DARPins. *Structure* 16: 1443-53
222. Derewenda ZS. 2010. Application of protein engineering to enhance crystallizability and improve crystal properties. *Acta Crystallogr D Biol Crystallogr* 66: 604-15
223. Dong A, Xu X, Edwards AM, Chang C, Chruszcz M, Cuff M, Cymborowski M, Di Leo R, Egorova O, Evdokimova E, Filippova E, Gu J, Guthrie J, Ignatchenko A, Joachimiak A, Klostermann N, Kim Y, Korniyenko Y, Minor W, Que Q, Savchenko A, Skarina T, Tan K, Yakunin A, Yee A, Yim V, Zhang R, Zheng H, Akutsu M, Arrowsmith C, Avvakumov GV, Bochkarev A, Dahlgren LG, Dhe-Paganon S, Dimov S, Dombrovski L, Finerty P, Jr., Flodin S, Flores A, Graslund S, Hammerstrom M, Herman MD, Hong BS, Hui R, Johansson I, Liu Y, Nilsson M, Nedyalkova L, Nordlund P, Nyman T, Min J, Ouyang H, Park HW, Qi C, Rabeh W, Shen L, Shen Y, Sukumard D, Tempel W, Tong Y, Tresagues L, Vedadi M, Walker JR, Weigelt J, Welin M, Wu H, Xiao T, Zeng H, Zhu H. 2007. In situ proteolysis for protein crystallization and structure determination. *Nat Methods* 4: 1019-21
224. Dyson MR, Perera RL, Shadbolt SP, Biderman L, Bromek K, Murzina NV, McCafferty J. 2008. Identification of soluble protein fragments by gene fragmentation and genetic selection. *Nucleic Acids Res* 36: e51
225. Mooij WT, Mitsiki E, Perrakis A. 2009. ProteinCCD: enabling the design of protein truncation constructs for expression and crystallization experiments. *Nucleic Acids Res* 37: W402-5
226. An Y, Yumerefendi H, Mas PJ, Chesneau A, Hart DJ. 2011. ORF-selector ESPRIT: a second generation library screen for soluble protein expression employing precise open reading frame selection. *J Struct Biol* 175: 189-97
227. Yumerefendi H, Tarendeau F, Mas PJ, Hart DJ. 2010. ESPRIT: an automated, library-based method for mapping and soluble expression of protein domains from challenging targets. *J Struct Biol* 172: 66-74
228. Mark BL, Mahuran DJ, Cherney MM, Zhao D, Knapp S, James MN. 2003. Crystal structure of human beta-hexosaminidase B: understanding the molecular basis of Sandhoff and Tay-Sachs disease. *J Mol Biol* 327: 1093-109
229. Derewenda ZS, Vekilov PG. 2006. Entropy and surface engineering in protein crystallization. *Acta Crystallogr D Biol Crystallogr* 62: 116-24
230. Goldschmidt L, Cooper DR, Derewenda ZS, Eisenberg D. 2007. Toward rational protein crystallization: A Web server for the design of crystallizable protein variants. *Protein Sci* 16: 1569-76
231. Walter TS, Meier C, Assenberg R, Au KF, Ren J, Verma A, Nettleship JE, Owens RJ, Stuart DI, Grimes JM. 2006. Lysine methylation as a routine rescue strategy for protein crystallization. *Structure* 14: 1617-22
232. Wukovitz SW, Yeates TO. 1995. Why protein crystals favour some space-groups over others. *Nat Struct Biol* 2: 1062-7
233. Banatao DR, Cascio D, Crowley CS, Fleissner MR, Tienson HL, Yeates TO. 2006. An approach to crystallizing proteins by synthetic symmetrization. *Proc Natl Acad Sci U S A* 103: 16230-5

234. Forse GJ, Ram N, Banatao DR, Cascio D, Sawaya MR, Klock HE, Lesley SA, Yeates TO. 2011. Synthetic symmetrization in the crystallization and structure determination of CelA from Thermotoga maritima. *Protein Sci* 20: 168-78
235. Yamada H, Tamada T, Kosaka M, Miyata K, Fujiki S, Tano M, Moriya M, Yamanishi M, Honjo E, Tada H, Ino T, Yamaguchi H, Futami J, Seno M, Nomoto T, Hirata T, Yoshimura M, Kuroki R. 2007. 'Crystal lattice engineering,' an approach to engineer protein crystal contacts by creating intermolecular symmetry: crystallization and structure determination of a mutant human RNase 1 with a hydrophobic interface of leucines. *Protein Sci* 16: 1389-97
236. Laganowsky A, Zhao M, Soriaga AB, Sawaya MR, Cascio D, Yeates TO. 2011. An approach to crystallizing proteins by metal-mediated synthetic symmetrization. *Protein Sci* 20: 1876-90
237. Kobe B, Center RJ, Kemp BE, Poumbourios P. 1999. Crystal structure of human T cell leukemia virus type 1 gp21 ectodomain crystallized as a maltose-binding protein chimera reveals structural evolution of retroviral transmembrane proteins. *Proc Natl Acad Sci U S A* 96: 4319-24
238. Liu Y, Manna A, Li R, Martin WE, Murphy RC, Cheung AL, Zhang G. 2001. Crystal structure of the SarR protein from Staphylococcus aureus. *Proc Natl Acad Sci U S A* 98: 6877-82
239. Song JJ, Liu J, Tolia NH, Schneiderman J, Smith SK, Martienssen RA, Hannon GJ, Joshua-Tor L. 2003. The crystal structure of the Argonaute2 PAZ domain reveals an RNA binding motif in RNAi effector complexes. *Nat Struct Biol* 10: 1026-32
240. Pioszak AA, Xu HE. 2008. Molecular recognition of parathyroid hormone by its G protein-coupled receptor. *Proc Natl Acad Sci U S A* 105: 5034-9
241. Bethea HN, Xu D, Liu J, Pedersen LC. 2008. Redirecting the substrate specificity of heparan sulfate 2-O-sulfotransferase by structurally guided mutagenesis. *Proc Natl Acad Sci U S A* 105: 18724-9
242. Ullah H, Scappini EL, Moon AF, Williams LV, Armstrong DL, Pedersen LC. 2008. Structure of a signal transduction regulator, RACK1, from Arabidopsis thaliana. *Protein Sci* 17: 1771-80
243. Mueller GA, Edwards LL, Aloor JJ, Fessler MB, Glesner J, Pomes A, Chapman MD, London RE, Pedersen LC. 2010. The structure of the dust mite allergen Der p 7 reveals similarities to innate immune proteins. *J Allergy Clin Immunol* 125: 909-17 e4
244. Zhan Y, Song X, Zhou GW. 2001. Structural analysis of regulatory protein domains using GST-fusion proteins. *Gene* 281: 1-9
245. Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SG, Thian FS, Kobilka TS, Choi HJ, Kuhn P, Weis WI, Kobilka BK, Stevens RC. 2007. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science* 318: 1258-65
246. Rosenbaum DM, Cherezov V, Hanson MA, Rasmussen SG, Thian FS, Kobilka TS, Choi HJ, Yao XJ, Weis WI, Stevens RC, Kobilka BK. 2007. GPCR engineering yields high-resolution structural insights into beta2-adrenergic receptor function. *Science* 318: 1266-73
247. Chun E, Thompson AA, Liu W, Roth CB, Griffith MT, Katritch V, Kunken J, Xu F, Cherezov V, Hanson MA, Stevens RC. 2012. Fusion partner toolchest for the stabilization and crystallization of G protein-coupled receptors. *Structure* 20: 967-76
248. Liu W, Chun E, Thompson AA, Chubukov P, Xu F, Katritch V, Han GW, Roth CB, Heitman LH, AP IJ, Cherezov V, Stevens RC. 2012. Structural basis for allosteric regulation of GPCRs by sodium ions. *Science* 337: 232-6
249. Koide S. 2009. Engineering of recombinant crystallization chaperones. *Curr Opin Struct Biol* 19: 449-57
250. Ostermeier C, Iwata S, Ludwig B, Michel H. 1995. Fv fragment-mediated crystallization of the membrane protein bacterial cytochrome c oxidase. *Nat Struct Biol* 2: 842-6

251. Lieberman RL, Culver JA, Entzminger KC, Pai JC, Maynard JA. 2011. Crystallization chaperone strategies for membrane proteins. *Methods* 55: 293-302
252. Krishnamurthy H, Gouaux E. 2012. X-ray structures of LeuT in substrate-free outward-open and apo inward-open states. *Nature* 481: 469-74
253. Li H, Dunn JJ, Luft BJ, Lawson CL. 1997. Crystal structure of Lyme disease antigen outer surface protein A complexed with an Fab. *Proc Natl Acad Sci U S A* 94: 3584-9
254. Domanska K, Vanderhaegen S, Srinivasan V, Pardon E, Dupeux F, Marquez JA, Giorgetti S, Stoppini M, Wyns L, Bellotti V, Steyaert J. 2011. Atomic structure of a nanobody-trapped domain-swapped dimer of an amyloidogenic beta2-microglobulin variant. *Proc Natl Acad Sci U S A* 108: 1314-9
255. Rasmussen SG, Choi HJ, Fung JJ, Pardon E, Casarosa P, Chae PS, Devree BT, Rosenbaum DM, Thian FS, Kobilka TS, Schnapp A, Konetzki I, Sunahara RK, Gellman SH, Pautsch A, Steyaert J, Weis WI, Kobilka BK. 2011. Structure of a nanobody-stabilized active state of the beta(2) adrenoceptor. *Nature* 469: 175-80
256. Tereshko V, Uysal S, Koide A, Margalef K, Koide S, Kossiakoff AA. 2008. Toward chaperone-assisted crystallography: protein engineering enhancement of crystal packing and X-ray phasing capabilities of a camelid single-domain antibody (VHH) scaffold. *Protein Sci* 17: 1175-87
257. Pardon E, Laeremans T, Triest S, Rasmussen SG, Wohlkonig A, Ruf A, Muyldermans S, Hol WG, Kobilka BK, Steyaert J. 2014. A general protocol for the generation of Nanobodies for structural biology. *Nat Protoc* 9: 674-93
258. Kohl A, Amstutz P, Parizek P, Binz HK, Briand C, Capitani G, Forrer P, Pluckthun A, Grutter MG. 2005. Allosteric inhibition of aminoglycoside phosphotransferase by a designed ankyrin repeat protein. *Structure* 13: 1131-41
259. Bandeiras TM, Hillig RC, Matias PM, Eberspaecher U, Fanghanel J, Thomaz M, Miranda S, Crusius K, Putter V, Amstutz P, Gulotti-Georgieva M, Binz HK, Holz C, Schmitz AA, Lang C, Donner P, Egner U, Carrondo MA, Muller-Tiemann B. 2008. Structure of wild-type Plk-1 kinase domain in complex with a selective DARPin. *Acta Crystallogr D Biol Crystallogr* 64: 339-53
260. Schweizer A, Roschitzki-Voser H, Amstutz P, Briand C, Gulotti-Georgieva M, Prenosil E, Binz HK, Capitani G, Baici A, Pluckthun A, Grutter MG. 2007. Inhibition of caspase-2 by a designed ankyrin repeat protein: specificity, structure, and inhibition mechanism. *Structure* 15: 625-36
261. Veesler D, Dreier B, Blangy S, Lichiere J, Tremblay D, Moineau S, Spinelli S, Tegoni M, Pluckthun A, Campanacci V, Cambillau C. 2009. Crystal structure and function of a DARPin neutralizing inhibitor of lactococcal phage TP901-1: comparison of DARPin and camelid VHH binding mode. *J Biol Chem* 284: 30718-26
262. Sennhauser G, Amstutz P, Briand C, Storchenegger O, Grutter MG. 2007. Drug export pathway of multidrug exporter AcrB revealed by DARPin inhibitors. *PLoS Biol* 5: e7
263. Kummer L, Parizek P, Rube P, Millgramm B, Prinz A, Mittl PR, Kaufholz M, Zimmermann B, Herberg FW, Pluckthun A. 2012. Structural and functional analysis of phosphorylation-specific binders of the kinase ERK from designed ankyrin repeat protein libraries. *Proc Natl Acad Sci U S A* 109: E2248-57
264. Nguyen HB, Hung LW, Yeates TO, Terwilliger TC, Waldo GS. 2013. Split green fluorescent protein as a modular binding partner for protein crystallization. *Acta Crystallogr D Biol Crystallogr* 69: 2513-23
265. Inokuma Y, Yoshioka S, Ariyoshi J, Arai T, Hitora Y, Takada K, Matsunaga S, Rissanen K, Fujita M. 2013. X-ray analysis on the nanogram to microgram scale using porous complexes. *Nature* 495: 461-6
266. Chaffotte AF, Guillou Y, Goldberg ME. 1992. Inclusion bodies of the thermophilic endoglucanase D from Clostridium thermocellum are made of native enzyme that resists 8 M urea. *Eur J Biochem* 205: 369-73

267. Gloster TM, Vocadlo DJ. 2012. Developing inhibitors of glycan processing enzymes as tools for enabling glycobiology. *Nat Chem Biol* 8: 683-94
268. Bischoff H. 1995. The mechanism of alpha-glucosidase inhibition in the management of diabetes. *Clin Invest Med* 18: 303-11
269. Hruska KS, LaMarca ME, Scott CR, Sidransky E. 2008. Gaucher disease: mutation and polymorphism spectrum in the glucocerebrosidase gene (GBA). *Hum Mutat* 29: 567-83
270. Spearman MA, Ballon BC, Gerrard JM, Greenberg AH, Wright JA. 1991. The inhibition of platelet aggregation of metastatic H-ras-transformed 10T1/2 fibroblasts with castanospermine, an N-linked glycoprotein processing inhibitor. *Cancer Lett* 60: 185-91
271. Moorthy NS, Ramos MJ, Fernandes PA. 2012. Studies on alpha-glucosidase inhibitors development: magic molecules for the treatment of carbohydrate mediated diseases. *Mini Rev Med Chem* 12: 713-20
272. Mechaly AE, Sassoon N, Betton JM, Alzari PM. 2014. Segmental helical motions and dynamical asymmetry modulate histidine kinase autophosphorylation. *PLoS Biol* 12: e1001776
273. Arai R, Ueda H, Kitayama A, Kamiya N, Nagamune T. 2001. Design of the linkers which effectively separate domains of a bifunctional fusion protein. *Protein Eng* 14: 529-32
274. Pavelka A, Chovancova E, Damborsky J. 2009. HotSpot Wizard: a web server for identification of hot spots in protein engineering. *Nucleic Acids Res* 37: W376-83
275. Ahmad S, Gromiha M, Fawareh H, Sarai A. 2004. ASAView: database and tool for solvent accessibility representation in proteins. *BMC Bioinformatics* 5: 51
276. Gibson DG. 2011. Enzymatic assembly of overlapping DNA fragments. *Methods Enzymol* 498: 349-61
277. An Y, Ji J, Wu W, Lv A, Huang R, Wei Y. 2005. A rapid and efficient method for multiple-site mutagenesis with a modified overlap extension PCR. *Appl Microbiol Biotechnol* 68: 774-8
278. Xiao YH, Pei Y. 2011. Asymmetric overlap extension PCR method for site-directed mutagenesis. *Methods Mol Biol* 687: 277-82
279. Hoover D. 2012. Using DNAWorks in designing oligonucleotides for PCR-based gene synthesis. *Methods Mol Biol* 852: 215-23
280. Hoover DM, Lubkowski J. 2002. DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Res* 30: e43
281. Grimm S, Yu F, Nygren PA. 2011. Ribosome display selection of a murine IgG(1) Fab binding affibody molecule allowing species selective recovery of monoclonal antibodies. *Mol Biotechnol* 48: 263-76
282. Kabsch W. 2010. Xds. *Acta Crystallogr D Biol Crystallogr* 66: 125-32
283. Evans PR. 2011. An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Crystallogr D Biol Crystallogr* 67: 282-92
284. McCoy AJ. 2007. Solving structures of protein complexes by molecular replacement with Phaser. *Acta Crystallogr D Biol Crystallogr* 63: 32-41
285. Murshudov GN, Vagin AA, Dodson EJ. 1997. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr* 53: 240-55
286. Terwilliger TC. 2004. Using prime-and-switch phasing to reduce model bias in molecular replacement. *Acta Crystallogr D Biol Crystallogr* 60: 2144-9
287. DeLano WL. 2002. The PyMOL Molecular Graphics System.
288. Kim Y, Bigelow L, Borovilos M, Dementieva I, Duggan E, Eschenfeldt W, Hatzos C, Joachimiak G, Li H, Maltseva N, Mulligan R, Quartey P, Sather A, Stols L, Volkart L, Wu R, Zhou M, Joachimiak A. 2008. Chapter 3. High-throughput protein purification for x-ray crystallography and NMR. *Adv Protein Chem Struct Biol* 75: 85-105
289. Esposito D, Garvey LA, Chakiath CS. 2009. Gateway cloning for protein expression. *Methods Mol Biol* 498: 31-54

290. Zhu B, Cai G, Hall EO, Freeman GJ. 2007. In-fusion assembly: seamless engineering of multidomain fusion proteins, modular vectors, and mutations. *Biotechniques* 43: 354-9
291. Berrow NS, Alderton D, Sainsbury S, Nettleship J, Assenberg R, Rahman N, Stuart DI, Owens RJ. 2007. A versatile ligation-independent cloning method suitable for high-throughput expression screening applications. *Nucleic Acids Res* 35: e45
292. Aslanidis C, de Jong PJ. 1990. Ligation-independent cloning of PCR products (LIC-PCR). *Nucleic Acids Res* 18: 6069-74
293. Warren TD, Coolbaugh MJ, Wood DW. 2013. Ligation-independent cloning and self-cleaving intein as a tool for high-throughput protein purification. *Protein Expr Purif* 91: 169-74
294. Bond SR, Naus CC. 2012. RF-Cloning.org: an online tool for the design of restriction-free cloning projects. *Nucleic Acids Res* 40: W209-13
295. Cabrita LD, Dai W, Bottomley SP. 2006. A family of E. coli expression vectors for laboratory scale and high throughput soluble protein production. *BMC Biotechnol* 6: 12
296. Luna-Vargas MP, Christodoulou E, Alfieri A, van Dijk WJ, Stadnik M, Hibbert RG, Sahtoe DD, Clerici M, Marco VD, Littler D, Celie PH, Sixma TK, Perrakis A. 2011. Enabling high-throughput ligation-independent cloning and protein expression for the family of ubiquitin specific proteases. *J Struct Biol* 175: 113-9
297. Kapust RB, Tozser J, Copeland TD, Waugh DS. 2002. The P1' specificity of tobacco etch virus protease. *Biochem Biophys Res Commun* 294: 949-55
298. Vincentelli R, Canaan S, Offant J, Cambillau C, Bignon C. 2005. Automated expression and solubility screening of His-tagged proteins in 96-well format. *Anal Biochem* 346: 77-84
299. Steen J, Uhlen M, Hober S, Ottosson J. 2006. High-throughput protein purification using an automated set-up for high-yield affinity chromatography. *Protein Expr Purif* 46: 173-8
300. DelProposto J, Majmudar CY, Smith JL, Brown WC. 2009. Mocr: a novel fusion tag for enhancing solubility that is compatible with structural biology applications. *Protein Expr Purif* 63: 40-9
301. Chatterjee DK, Esposito D. 2006. Enhanced soluble protein expression using two new fusion tags. *Protein Expr Purif* 46: 122-9
302. Cheng Y, Gu J, Wang HG, Yu S, Liu YQ, Ning YL, Zou QM, Yu XJ, Mao XH. 2010. EspA is a novel fusion partner for expression of foreign proteins in Escherichia coli. *J Biotechnol* 150: 380-8
303. Bernard P, Gabant P, Bahassi EM, Couturier M. 1994. Positive-selection vectors using the F plasmid ccdB killer gene. *Gene* 148: 71-4
304. Quan S, Koldewey P, Tapley T, Kirsch N, Ruane KM, Pfizenmaier J, Shi R, Hofmann S, Foit L, Ren G, Jakob U, Xu Z, Cygler M, Bardwell JC. 2011. Genetic selection designed to stabilize proteins uncovers a chaperone called Spy. *Nat Struct Mol Biol* 18: 262-9
305. Scolnik PA. 2009. mAbs: a business perspective. *MAbs* 1: 179-84
306. Vasella A, Davies GJ, Bohm M. 2002. Glycosidase mechanisms. *Curr Opin Chem Biol* 6: 619-29
307. Jones S, Thornton JM. 1996. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A* 93: 13-20
308. Kwong PD, Wyatt R, Robinson J, Sweet RW, Sodroski J, Hendrickson WA. 1998. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature* 393: 648-59
309. Kwong PD, Wyatt R, Desjardins E, Robinson J, Culp JS, Hellmig BD, Sweet RW, Sodroski J, Hendrickson WA. 1999. Probability analysis of variational crystallization and its application to gp120, the exterior envelope glycoprotein of type 1 human immunodeficiency virus (HIV-1). *J Biol Chem* 274: 4115-23

310. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. 2003. Protein disorder prediction: implications for structural proteomics. *Structure* 11: 1453-9
311. Reetz MT, Kahakeaw D, Lohmer R. 2008. Addressing the numbers problem in directed evolution. *Chembiochem* 9: 1797-804
312. Koide A, Gilbreth RN, Esaki K, Tereshko V, Koide S. 2007. High-affinity single-domain binding proteins with a binary-code interface. *Proc Natl Acad Sci U S A* 104: 6632-7

ANEXO I:

Tuning different expression parameters to achieve soluble recombinant proteins in E. coli: Advantages of high-throughput screening.

Agustín Correa[1] and Pablo Oppezzo[1]

[1]Recombinant Protein Unit, Institut Pasteur de Montevideo, Uruguay

Biotechnol J. 2011 Jun;6(6):715-30.

Review

Tuning different expression parameters to achieve soluble recombinant proteins in *E. coli*: Advantages of high-throughput screening

Agustín Correa and Pablo Oppezzo

Recombinant Protein Unit, Institut Pasteur de Montevideo, Uruguay

Proteins are the main reagents for structural, biomedical, and biotechnological studies; however, some important challenges remain concerning protein solubility and stability. Numerous strategies have been developed, with some success, to mitigate these challenges, but a universal strategy is still elusive. Currently, researchers face a plethora of alternatives for the expression of the target protein, which generates a great diversity of conditions to be evaluated. Among these, different promoter strength, diverse expression host and constructs, or special culture conditions have an important role in protein solubility. With the arrival of automated high-throughput screening (HTS) systems, the evaluation of hundreds of different conditions within reasonable cost and time limits is possible. This technology increases the chances to obtain the target protein in a pure, soluble, and stable state. This review focuses on some of the most commonly used strategies for the expression of recombinant proteins in the enterobacterium *Escherichia coli*, including the use of HTS for the production of soluble proteins.

Received 1 February 2011
Revised 15 March 2011
Accepted 21 March 2011

Supporting information available online

Keywords: Directed evolution· High-throughout screening · Protein folding · Recombinant protein · Soluble protein expression

1 Introduction

Purified and soluble proteins are essential tools in academic and medical areas. The knowledge of the molecular structure of individual proteins allows important questions about the physiological function of these molecules to be addressed, so as to know the biochemical and regulator pathways in which they are implicated. The pharmaceutical industry and biotechnology laboratories are primarily interested in the development of recombinant proteins (RPs). Obtaining purified and functional proteins is a key issue for modern biotechnology laboratories.

Natural protein sources rarely meet the requirements for quantity and ease of isolation; hence, recombinant technology is often the method of choice. Recombinant cell factories are constantly employed for the production of protein preparations bound for downstream purification and processing. In the 1980s, the development of genetic engineering made the production and expression of target proteins in a recombinant form possible by using different expression hosts, including bacterial, fungal, or eukaryotic host cells.

In all of these expression systems, the use of the enterobacterium *Escherichia coli* is the most widely used. This microbial factory was the first host to be used for RP production almost 40 years ago [1], and until now, approximately 60% of all RPs reported in the literature were expressed using *E. coli*. [2]. The main reasons for the extensive use of this bac-

Correspondence: Dr. Oppezzo Pablo, Institut Pasteur de Montevideo, Unit of Recombinant Protein, Mataojo 2020, Montevideo (11400), Uruguay
E-mail: poppezzo@pasteur.edu.uy

Abbreviations: **GST**, glutathione-S-transferase; **HTS**, high-throughput screening; **IMAC**, immobilized metal ion affinity chromatography; **IPTG**, isopropyl-β-D-thiogalactopyranoside; **MBP**, maltose binding protein; **NusA**, N-utilization substance A; **RP**, recombinant protein; **SUMO**, small ubiquitin-like modifier protein; **trx**, thioredoxin

terium in this area are as follows: extensive knowledge of the genetics of the bacterium (large number of cloning vectors and mutant host strains commercially available), ease of use, low cost, and a high yield of the target protein [3, 4].

The use of *E. coli*, however, for RP production has encountered several disadvantages. For example, many of the post-translational modifications found in eukaryotes, such as N- and O-glycosylation, amidation, hydroxylation, myristoylation, palmitation, or sulfation, are absent in *E. coli* [5], which limits its application. On top of this, the high expression levels of RP can often lead to the accumulation of aggregated insoluble protein, resulting in inclusion-body formation in the cytoplasm of the bacteria [6]. High translation rate can be a serious problem when the target protein is a heterologous molecule. Thus, the soluble expression and native purification of the target protein in *E. coli* remains an important bottleneck in the production area of RP. Nevertheless, if the protein to be expressed is cytoplasmic, lacks the above-mentioned post-translational modifications, possesses few disulfide bonds, and does not present a multidomain composition, the use of the *E. coli* as the host is the recommended choice for the first trials of protein production [7].

Production of RP in *E. coli*, whether for biochemical analysis, therapeutics, or structural studies, requires the success of mainly two crucial steps: (i) soluble expression of the target protein; and (ii) purification and stabilization of a functional molecule.

In the past three decades considerable efforts to improve the production of soluble and functional RP have been carried out. These advances include the development of different expression strains [8], a wide variety of plasmids under the control of different promoters, or the use of special tags [9]. The co-expression of target protein with molecular chaperones or folding modulators has also been employed [10, 11], as well as the introduction of mutations in the target gene [12]. Additionally, diverse growth temperatures, different induction densities, as well as changes in media composition are also important variables evaluated with the purpose of improving the solubility and purification of the target protein. Because soluble does not always mean functional, quite often the protein can form soluble aggregates that can be unfolded, may be inactive, and/or difficult to crystallize, making the soluble protein useless. Therefore, it is also important to characterize the aggregation state of the protein after expressing the target protein. In this regard, the use of analytical gel filtration [13, 14] and/or static or dynamic light scattering [15–17] could be used with this purpose.

This review focuses on the solubility problem of RP in *E. coli*, linking two principal issues: (i) the most useful and general strategies employed for the expression of RP in this bacteria; and (ii) the use of high-throughput screening (HTS) techniques to find the optimal parameters to obtain soluble RPs.

2 Selecting a vector for RP expression

Selection of the vector is one of the first issues that the researcher faces when trying to express a RP. Vector characteristics will affect many important variables essential for the success of protein production: (i) localization of the target protein in the bacterial microenvironment; (ii) plasmid copy number as a consequence of the replication origin; (iii) promoter type, which modulates the protein yield, the rate of transcription, and the stringency of repression before induction; (iv) fused proteins and/or fused tags, which could influence protein solubility and/or stability; and (v) co-expression of the target protein with molecular partners or chaperones able to help in the folding process.

2.1 Localization of target protein

A limitation of the production of properly folded proteins in *E. coli* has been the relatively high reducing potential of the cytoplasmic compartment: disulfide bonds are usually formed only upon export into the periplasmic space [18–20]. Most often RPs are expressed in the cytoplasm of *E. coli*; however, when the RP needs the presence of disulfide bonds one option is to perform the expression in the *E. coli* periplasmic space.

In this context, many vectors have been modified to export the protein target into the periplasm. For this purpose, vectors carrying signal peptides (sequence for periplasmic export) are commercially available. Expression systems such as pMalp2 (New England BioLabs) or pET 22b (Novagen) are normally used. The principal drawback of this strategy is that the translocation machinery to the periplasm of *E. coli* could be easily saturated, decreasing the final yield of RP [8].

2.2 Plasmid copy number

The origin of replication is responsible for the plasmid copy number and determines the gene dosage accessible for protein expression. The copy number for common *E. coli* expression plasmids ranges

from low (2 to 20) to high (20 to 40) [21]. Usually a high copy number is desired for maximum gene expression, but in some instances this can lead to a metabolic burden that negatively affects the biomass and the final yield [22, 23]. Among the most commonly used origins of replication in plasmids for RP expression are the ColE1 (high copy) and the p15A (low copy) [24]. Replication origins are important when co-expressing proteins from different plasmids. In these cases, each vector may contain a different origin of replication because plasmids with the same origin are mutually exclusive in the same bacterial host [25].

2.3 Promoter type

The promoter sequence is a central element that affects the strength and duration of transcription, and in turn, protein yield. Recombinant expression needs to be strong, present a very low basal expression level, and its induction should be simple and cost effective. Along with inducers, they can be thermal and chemical, of which the chemical inducer isopropyl-β-D-thiogalactopyranoside (IPTG; a nonhydrolyzable lactose analogue) is the most commonly used [16, 21, 26]. In this section, we discuss the promoters frequently used for protein expression, such as T7 (Novagen); T5 (Quiagen) and the hybrid promoters, such as *trc* and *tac* (Invitrogen and Sigma, respectively); pBAD promoter (Invitrogen); and finally temperature-controlled promoters, such as CspA and the phage promoters p_L/p_R.

2.3.1 T7 promoter

The T7-based pET expression system is by far the most used system for recombinant expression in *E. coli* [16, 27, 28]. It was first described by Studier and co-workers [29, 30] and is based in the highly selective T7 RNA polymerase from phage T7 to drive RP production. This polymerase only transcribes genes under the control of the T7 promoter and it has been shown that it can transcribe eight times faster than *E. coli* polymerases, producing a high yield of protein [31]. The T7 promoter is considered a strong promoter and RP could reach up to 50% of the total cell proteins. Because *E. coli* lacks this polymerase, some strains, such as BL21(DE3), have been developed that contain a chromosomal copy of the T7 polymerase gene, under the control of the lac promoter derivative lacUV5 [29, 32]. The lacUV5 promoter contains point mutations that increase the promoter strength and make it less sensitive to catabolite repression [33]. In this way, the promoter is only controlled by the lac repressor, LacI, which allows induction with IPTG, even in the presence of glucose. The addition of IPTG releases the repression caused by the binding of LacI to the lac operator, resulting in the expression of T7 polymerase, which in turn transcribes the target gene with the concomitant production of the RP [34].

2.3.2 T5 and hybrid promoters

The essential element of this unit is a promoter derived from coliphage T5 that is utilized by *E. coli* RNA polymerase. This promoter system has been used mainly in pQE vectors (QIAGEN) combined with a double *lac* operator repression module to provide tightly regulated level expression of RPs in *E. coli*. Protein synthesis is induced with IPTG, but, in contrast to the T7 promoter system, is more effectively blocked in the presence of high levels of *lac* repressor with higher stability of the cytotoxic constructs as a consequence [35].

The trc and tac promoters are hybrids of naturally occurring promoters, consisting of the –35 region of *trp* promoter and the –10 region of *lacUV5* promoter [36, 37]. The expression is also induced by IPTG and although they are considered as strong promoters, their production is lower than that of T7 promoters (about 15–30 % of the total cellular protein) [38].

2.3.3 araBAD promoter

Another promoter system is the *araBAD* (P_{BAD}) promoter of the arabinose operon. When a gene is cloned downstream of the P_{BAD} promoter, the expression is regulated by the *araC* protein, which is a positive and negative regulator of the P_{BAD} promoter. In this system, induction is achieved by the addition of L-arabinose (usually 0.2% w/v) in a titratable manner, showing a linear increase of protein expression upon increasing inducer concentration. Similar to the T5 promoter system, P_{BAD} is a tightly regulated promoter, making it ideal for the expression of highly toxic proteins [39, 40]. Moreover, basal expression levels can be reduced even more by the addition of glucose or the anti-inducer fucose, which represses expression [41]. Compared with the T7 promoter system, in some cases it has been observed that P_{BAD} results in lower yields [21, 42].

2.3.4 Temperature-controlled promoters

Instead of using a chemical inducer, some promoters are induced upon a physical signal, such as a decrease or increase in temperature. CspA protein is the major cold shock protein of *E. coli* and is virtually undetectable at 37°C, but after transferring the culture to 15°C, the production of this protein could be greater than 10% [43]. Therefore, genes under the control of the *cspA* promoter can be in-

duced simply by a downshift in temperature between 15 and 25°C [44–46]. This expression at low temperatures could be beneficial for the soluble expression of aggregation-prone proteins [44, 47]. A series of vectors were developed that contain the *lac* operator sequence immediately upstream of the *cspA* transcription initiation site to prevent leaky expression from *cspA* promoter at 37°C [48]. In this case, induction is achieved by temperature downshift and the addition of IPTG.

Other promoters, such as pL/pR phage lambda promoter, are induced after increasing the culture temperature. In this system, the pL (leftward) and pR (rightward) strong promoters are regulated by the temperature sensitive mutant cI857 repressor of bacteriophage λ [49]. At low temperatures (usually 28–32°C), transcription is inhibited by the binding of cI857 to the pL or pR promoters. After increasing the temperature above 37°C (usually 40–42°C), cI857 binding is released from the promoter and gene expression is induced [49, 50, 51].

2.4 Fused proteins and/or fused tags

Fused proteins and/or fused tags has been widely used with the aim of solving the two main obstacles in the expression of RP field: solubility and stability of target protein before purification [9]. With the advent of HTS the use of fused tags has become an essential tool to permit the use of a generic purification strategy.

The most common and commercially available short tags are the histidine (his-tag) and strep tags, whereas complete proteins used as fusion tags are glutathione-S-transferase (GST), maltose-binding protein (MBP), thioredoxin (trx), and more recently the small ubiquitin-like modifier protein (SUMO) and N-utilization substance A (NusA).

Short tags can be fused to the N and/or C terminus of the target protein, whereas complete proteins are usually placed at the N terminus of the RP to not only aid the purification step but also to improve the solubility. In both cases, a short flexible hydrophilic linker is usually placed between the tag and the target protein to allow the accessibility of the tag in the purification step and to introduce a specific cleavage site for its removal [9].

Short affinity tags dedicated to isolate the target protein include those given in the following sections.

2.4.1 His-tag

The his-tag generally consists of six histidine residues in tandem (0.84 kDa) and exploits the capacity of this residue to reversibly interact with metal ions immobilized in a metal-chelate matrix [52]. Immobilized metal ion affinity chromatography (IMAC), is the most widely used method for purifying RP. The Ni^{2+} metal ion (GE, QUIAGEN) is commonly used for purification, but other transition metals, such as Co^{2+} (TALON, Clontech), Cu^{2+}, and Zn^{2+}, have also been used with success. Because the tertiary structure of the His-tag is not important for coordination with the metal, His-tagged RP can be purified by using IMAC under denaturing conditions and subsequently the target protein is refolded [53]. Once immobilized, the RP can be eluted from the matrix by the addition of imidazole (up to 0.5M), or by lowering the pH (pH < 5).

Nevertheless, the use of imidazole is by far the most commonly used method because it is milder and allows the use of a fine gradient to improve protein purity without affecting RP stability [16]. The IMAC purification procedure has been fully automated and the vast majority of structural genomic centers use it as their main affinity strategy. Automation has been achieved at the microscale level for searching for soluble-protein expression by magnetic beads or filtration-based purification protocols in a 96-well format [28, 54–58]. These protocols allow the purification of hundreds of different conditions per week. On a larger scale, the use of positive pressure for liquid transfer through different columns, permitted processing of up to 60 cell lysates in 18.5 h to give milligram yields of the target protein [59]. Finally, by using specific antibodies against the His-tag, the evaluation of different conditions can be easily made by dot blot [60].

One drawback of IMAC purification is its susceptibility to metal chelators. This was evidenced in a recent work, in which *E. coli* lysate severely reduced the binding capacity of the column [61]. This reduction was caused by low-molecular-weight components (such as metallophores) that are associated with the periplasmic space. This effect is more important when working with low-abundance RP and higher culture sizes are necessary to increase the target protein yield [61].

2.4.2 Strep-TagII (Strep-Tactin)

Strep-TagII (Strep-Tactin) is another attractive affinity tag formed by eight amino acids (WSHPQFEK) that binds in a reversible way to an engineering variant of streptavidin [62]. Like the his-tag, the strep-tagII is biologically inert, proteolytically stable, and does not interfere with protein folding. This highly specific system allows the isolation of the target protein in a pure state and elution of the protein is obtained by using mild buffer conditions by competition with D-biotin or preferentially D-desthiobiotin [62]. In a comparative study, it was shown that this tag had a better

cost–benefit relationship than other tags and was a very good compromise of high purity with good yields at a moderate cost [63].

2.4.3 Calmodulin-binding peptide, S-tag, and Si-tag

Other affinity tags include the calmodulin-binding peptide (26 amino acids), which binds specifically to calmodulin in a calcium-dependent manner, allowing proteins with this tag to be purified over calmodulin resin where the elution is done through the addition of a buffer containing 2 mM ethylene-glycol-tetra-acetic acid (EGTA) [64]. S-tag (15 amino acids), derived from the N-terminal helix of RNAse A, is another used tag, normally eluted with S-protein [65].

Recently an Si-tag was described by Ikeda et al [66]. This is a large tag (30 kDa) based on the reversible and specific binding of the bacterial ribosomal protein L2 to silica surfaces. After binding, the target protein can be eluted in a pure state from the silica by the addition of a buffer containing a high concentration of a divalent cation, such as 2M $MgCl_2$. Since silica serves as both a resin and ligand for Si-tag, this method is one of the cheapest for the isolation of tagged proteins [66].

Other tags involving complete proteins provide dual purposes: on the one hand, they allow a simple protein purification step and, on the other, they offer the possibility of improving the solubility of the target protein. As mentioned above, among the most widely used solubility-enhancer tags found are GST, MBP, Trx, SUMO, and NusA.

2.4.4 GST tag

The GST tag is normally used at the N terminus of target protein, binds tightly to glutathione resin, and can be eluted by the addition of reduced glutathione [67]. It is important to note that GST (26 kDa) dimerizes, thus it is not recommended for proteins that are prone to aggregation [68]. Several studies have shown that GST is a poor solubility enhancer [7, 69], but is still a widely used fusion tag and allows RPs to be purified in a single step [67, 70].

2.4.5 MBP tag

The MBP tag is a soluble periplasmic protein from *E. coli* that can bind strongly to amylose resins and the fusion protein can be eluted by the application of free maltose [71]. MBP has been shown to enhance the protein solubility when it is fused as both N- and C-terminal fusion tags [7, 72, 73]. It can also be used to target proteins to the periplasm if the endogenous signal sequence, *malE*, is included in the gene [21]. In a comparative study, it was found that MBP was more efficient in solubilizing the fusion partner than GST and thioredoxin [74]. One drawback of this tag is its large size (42 kDa), which can interfere with the biological activity of the RP.

2.4.6 Trx tag

The trx tag is another solubility tag derived from *E. coli trxA*. This protein (11.6 kDa) is an oxido-reductase that facilitates the reduction of other proteins and has some properties that make it suited as a fusion partner. When *trxA* is expressed in *E. coli*, it can accumulate in a fully soluble state of up to 40% of the total cellular protein [75]. This suggests that thioredoxin translates very efficiently, thus, if fused at the N terminus, this property can be conferred to the partner target protein [76, 77]. As well as this, it has been found that thioredoxin has a high thermal stability (*Tm*: 85°C), that can contribute to fusion partner stabilization [78].

2.4.7 SUMO and NusA tags

Other solubility-enhancer fusion tags that are gaining ground are SUMO (11.2 kDa) and NusA (55 kDa). Yeast SUMO (Smt3) is commonly fused to the N terminus of target proteins and can improve the solubility and expression of the fused protein. A comparative study showed the utility of SUMO as a fusion partner, in which it behaved better than other common tags, with the added advantage that it generated a native N terminus for the target protein after cleavage with a specific protease [79, 80]. Finally, NusA is a transcription elongation and anti-termination factor of *E. coli* [81], which, as a fusion protein, also showed improvement in the expression and solubility of target proteins when fused as an N-terminal tag [79, 82–84].

A drawback to these fusion-tag strategies that must be considered is the occurrence of false positives. In many constructs it has been commonly observed that a soluble fusion target protein became insoluble after cleavage with the specific endoprotease. This can imply that the fused protein is held in solution as a result of interactions between the solubility partner and not as a result of a native fold of the target protein [85].

2.5 Tag cleavage

Because all fusion tags can interfere with structural and functional studies of the expressed protein, it is usually necessary to remove the fused tag after purification of the target protein. This is done by the addition of a specific endoprotease cleavage site between the tag and the target protein, as mentioned earlier. Several specific proteases used for this task are commercially available (thrombin, factor Xa, and enterokinase). Another very specific

protease is the 3C-type protease from the tobacco etch virus (TEV) [21]. This molecule cleaves the sequence ENLYFQ/G specifically [86] and, relative to other proteases, does not present nonspecific secondary cleavage.

2.6 Co-expression of target protein with molecular partners or chaperones

In some instances, the expression of soluble RP is achieved by co-expressing a biological partner of the target protein or molecular chaperones that help with correct folding of the RP. Co-expression can be achieved using single or multiple expression vectors. The co-expression of a biological partner has been used mainly to improve the solubility of the target protein, but in some specific cases protein complex formation is also required for optimal activity [87–89]. In a different approach, it has been suggested that the production of slow-folding RP can overwhelm the host chaperones, leading to the accumulation of the target proteins as inclusion bodies [90]. Thus, supplementing with co-expression of molecular chaperones, such as the chaperone set DnaK/DnaJ/GrpE (KJE) or GroEL/GroES (ELS), ClpB, the small heat-shock proteins IbpA and IbpB, and the ribosome-associated trigger factor, minimized aggregation and improved the solubility of many RPs [21, 91, 92].

3 Performing new cloning strategies

As mentioned earlier, to find suitable conditions for the soluble expression of target proteins, different combinations of promoters and fusion tags need to be evaluated. This requires the cloning of the target gene in several plasmids, therefore, the use of a method that enables the easy transfer of the gene into multiple vectors regardless of the target sequence is preferable to a classical restriction strategy. Examples of this strategy include the commercially available Gateway® (Invitrogen) [93] and In-Fusion™ (Clontech) [94, 95] methods, which rely on a recombination process between the insert and the destination vector. Other ligation-independent cloning (LIC) methods based on the use of complementary single strands for the fusion of the insert within the vector are also used for the easy cloning of several genes in different vectors [96–100]. Recently, adaptation of an LIC strategy based on the integration of a target gene into an expression vector by whole-plasmid amplification of the plasmid and the insert was developed, known as RF cloning [101] (Fig. 1A). In this method, after amplification of target gene, the PCR product is used as a megaprimer in a second PCR to amplify the whole plasmid. The parental DNA is eliminated by cleavage with DpnI and the newly synthesized plasmid containing the insertion is used to transform *E. coli* cells [101]. The great advantage of this method is that it can be used with any destination vector, such as the commercially available pET, pQE, and pACYCDuet vectors, and the insertion can occur at any desired position without the addition of any unnecessary sequences to the target gene [101].

4 Selecting a host strain for RP expression

The selection of the *E. coli* strain can have a profound impact on the RP production, since it gives the genetic background in which protein synthesis occurs. Different commercial *E. coli* strains have been developed that facilitate the soluble production of heterologous proteins. The selection of the expression strain is based on the characteristics of the target protein, such as whether the protein contains disulfide bonds, is highly toxic, or contains rare codons caused by the heterologous taxonomic origin of the target protein. In this context, different strains could be grouped as described in the following sections.

4.1 Protease-deficient strains

The *E. coli* BL21 and its derivatives are most commonly used for RP expression. BL21 is deficient in the adenosine triphosphate (ATP)-dependent protease Lon [102] and in the outer-membrane protease OmpT [103], thus reducing the degradation of the target protein and improving the yield. The BL21(DE3) derivative is deficient in OmpT/Lon proteases and contains a chromosomal copy of the T7 RNA polymerase under the control of the lacUV5 promoter for the expression of RP under the control of the T7 promoter.

4.2 Stringent repression of RP expression

Because the robust transcription of the T7 polymerase, even its minimal basal production, results in a leaky expression of the target gene prior to induction. This could be detrimental for the host if the target protein is toxic or even prevent the establishment of the plasmid carrying the toxic gene [29]. To reduce this basal level of expression, several host strains have been developed that contain plasmid coding for the natural inhibitor of T7 polymerase, the bacteriophage T7 lysozyme [104]. Usually pLysS and pLysE plasmids are commercially available as BL21(DE3)pLysS and

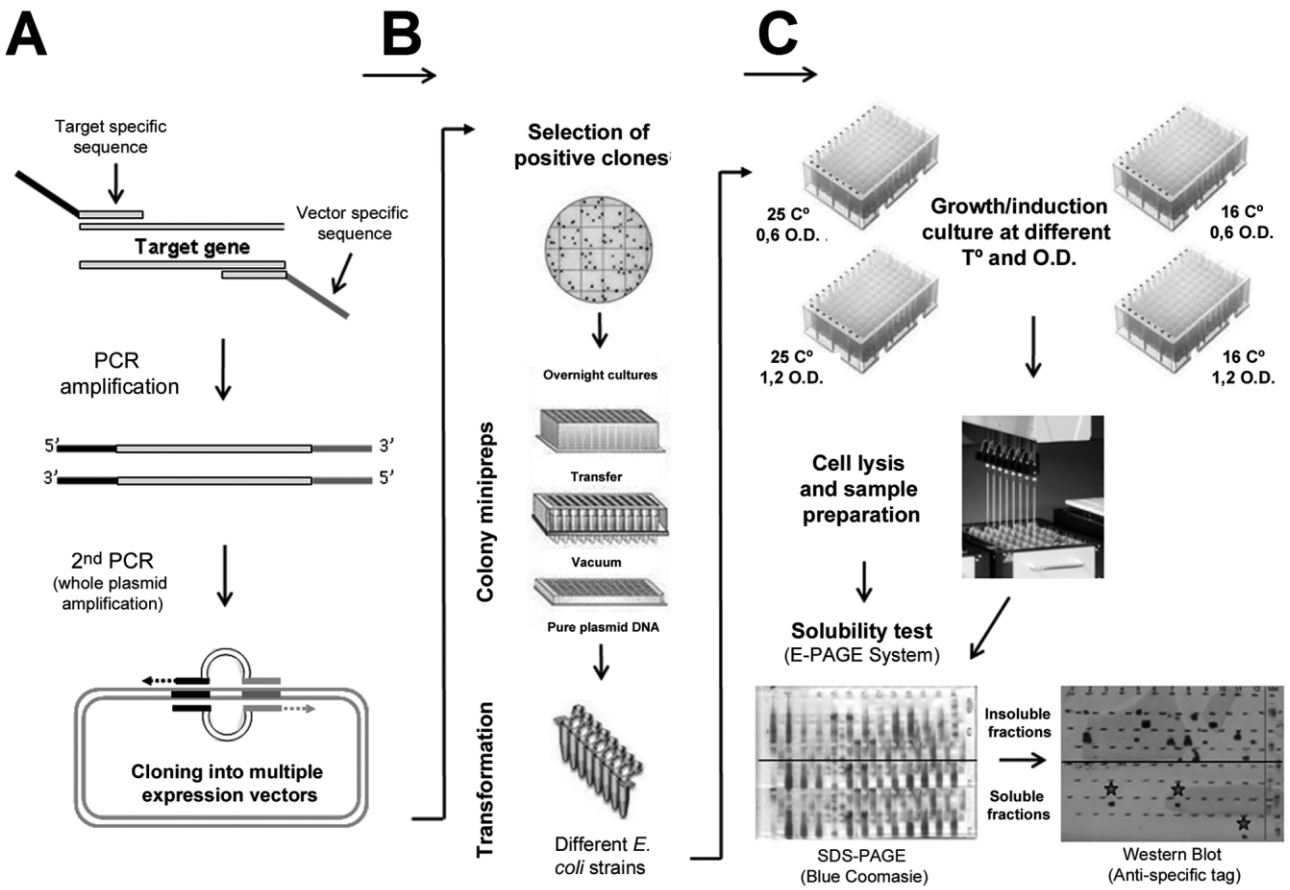


Figure 1. Schematic representation of automated HTS for the soluble expression of RP. (**A**) HTS of different constructs and details of the RF cloning methodology. This stage is manually performed in our system. (**B**) HTS of different expression strains; selection of positive clones is manually achieved, whereas colony minipreps and the transformation process is done in the robotic workstation. (**C**) HTS of different culture conditions; this module is completely achieved in our robotic platform. Finally, soluble protein conditions are manually determined by western blot analysis. The soluble expression conditions established for a specific target protein are indicated by the stars.

BL21(DE3)pLysE (Novagen). While pLysS plasmids produce low levels of T7 lysozyme, pLysE plasmids provide higher levels of the inhibitor [104]. Because T7 lysozyme continues to inhibit T7 polymerase after induction, this could result in lower yields of the target protein for nontoxic proteins. Another characteristic of T7 lysozyme is that it can cut a specific bond in the peptidoglycan layer of the *E. coli* cell wall, which can reduce the growth rate of strains harboring pLys plasmids, but have the benefit of facilitating cell lysis for protein purification [104]. An attractive alternative is to use an *E. coli* strain that contains the T7 RNA polymerase under the control of a more stringent promoter, such as the aforementioned P_{BAD}, instead of the lacUV5. Such is the case of the *E. coli* BL21AI strain commercialized by Invitrogen. In this case, induction is achieved by the addition of L-arabinose to a final concentration of 0.2% and, if working with expression vectors with the lacI gene, it is also necessary to add 1 mM IPTG. This strain has a four-fold lower basal expression level with similar yields of RP when compared with BL21(DE3)pLysS [39], making it a very suitable strain for the expression of highly toxic genes.

4.3 Expression of disulfide bond-containing proteins

One of the most common post-translational modifications is disulfide-bond formation. In this issue, Salinas et al. review the biochemical bases of this modification in more detail [105]. Disulfide-bond formation occurs in the periplasm of *E. coli*, which is a more oxidizing compartment than the cytoplasm, by Dsb proteins (DsbA, B, C, D and G). Whereas DsbA is responsible for disulfide-bond formation, the isomerases DsbC and DsbG are responsible for the rearrangement or isomerization of incorrectly formed bonds. Finally, DsbB and

DsbD membrane proteins recycle DsbA and DsbC/G, respectively [106]. One strategy to produce proteins with disulfide bonds is to direct RP expression to the periplasm with the addition of an N-terminal signal peptide. Another way is to change the redox state of the *E. coli* cytoplasm to a more oxidative environment. The reduced environment found in the cytoplasm of *E. coli* is actively maintained by the action of pathways involving the glutathione reductase (*gor*) and thioredoxin reductase (*trxB*) [107–109]. Therefore, single (*trxB*$^-$) and double (*trxB*$^-$/*gor*$^-$) mutants of *E. coli* have been developed, and commercialized by Novagen as AD494 and Origami, respectively, that enhance the production of disulfide-bond-containing proteins in the cytoplasm of *E. coli*. Another *E. coli* strain was developed that expresses DsbC in the cytoplasm and also contains the *trxB*$^-$/*gor*$^-$ mutations, thus improving the correct formation of disulfide bonds (SHuffle, by New England Biolabs).

Recently, it was possible to produce disulfide-bond-containing proteins in the cytoplasm of *E. coli* by the co-expression of a catalyst of disulfide-bond formation from *Saccharomyces cerevisiae,* Erv1p, without the need to disrupt the reducing pathways of the host (*trxB*$^-$ and/or *gor*$^-$ mutants) [110].

4.4 Expression of membrane proteins

More than a decade ago, two *E. coli* mutant host strains derived from BL21(DE3) were generated and isolated for the production of difficult-to-express proteins, such as membrane proteins. Named C41(DE3) and C43(DE3) [111], they are commercially available for the expression of toxic and membrane proteins (Lucigen). It was found that the reason for their improved over-expression of membrane proteins was the result of mutations in the lacUV5 promoter. These negatively affected the expression of the T7 polymerase, delaying the expression of the target protein, and preventing the saturation of membrane-translocation machinery. In addition to these mutations, C43(DE3) also slows down the expression of the lactose permease (LacY), delaying the expression of the target protein even more [8]. Overall, this work led to the development of a BL21(DE3) derivative, named Lemo21(DE3) (New England Biolabs), that contained a plasmid encoding for the T7 lysozyme under the control of the L-rhamnose-inducible promoter (*rha*BAD). This is a titratable promoter that allows the production of different levels of T7 lysozyme upon addition of different amounts of L-rhamnose (0–2000 μM) [8]. Therefore, by adding higher concentrations of L-rhamnose, more lysozyme is expressed and less T7 RNA polymerase is available, thus controlling the rate of transcription of the target protein.

4.5 Expression of codon-biased genes

When using *E. coli* as a host, some obstacles can be found as a result of codon biases. When the codon usage of the target gene differs from that of the expression host, the low-abundance tRNAs from the host are depleted by the rare codons present in the foreign mRNA and can result in amino acid misincorporation and/or truncation of the polypeptide chain, thus affecting heterologous gene expression [112]. One alternative to the aforementioned problem is to perform a rational gene optimization, by de novo gene synthesis, where the rare codons of the target gene are changed to codons more frequently used in *E. coli*. This methodology can be successful for many cases [113], but is also expensive. Another strategy is to use an *E. coli* host with supplemented low-abundance tRNAs, thus improving codon biases [112]. At present, numerous strains containing plasmid coding for rare tRNAs are commercially available (Rosetta2 and Rosetta2 (DE3) from Novagen and BL21-CodonPlus(DE3)-RIPL, RIL, and RP from Stratagene). However, for some genes, these low-abundance codons are necessary to provoke a pause in the ribosome processing, which allows the correct folding of an intermediate in the newly synthesized chain. In these cases, the use of a strain with supplemented codons could be detrimental to protein solubility [114].

5 Optimizing variables in *E. coli* growth: Temperature and media effects

Another common strategy to express the target protein in a soluble state is to evaluate different culture conditions, such as the growth temperature after induction as well as the composition of growth media.

Quite often, lowering temperature during induction (15–25°C) improves the solubility of the RP by diminishing aggregation and inclusion-body formation. This could be the result of a decrease in the rate of protein production, allowing the newly synthesized chain to fold properly [115]. Thus, it is highly recommended to evaluate different induction temperatures when searching for conditions for soluble protein expression [16]. Because one must test many conditions in parallel for protein expression, it is necessary to perform cell growth in many small cultures. This is usually done in multi-well plates and commonly used media include LB, 2YT, terrific broth, and minimal media (M9).

Growth in a multi-well plate format can result in a situation in which different cultures are ready for induction at different times due to differences in growth rates. This leads to a scenario in which induction is not homogeneously distributed among cultures, making the comparison between different conditions difficult.

One solution to such a problem is the use of a medium in which the uptake of the inducer depends on the metabolic state of the bacteria. In this regard, Studier developed an autoinduction medium in which the inducer, lactose (in systems regulated by the LacI), was prevented from inducing cells by compounds that could be depleted during growth (e.g., glucose) [116]. During the initial growth period, glucose was preferentially used as an energy and carbon source instead of lactose. After glucose was depleted, lactose and glycerol were metabolized and the target protein was induced automatically. In this way, there was no need to monitor bacterial growth and add inducer at the proper time, making it suitable for HTS [116]. Also, early basal expression was prevented by catabolite repression, making it suitable for the expression of toxic proteins. In this medium, glucose, glycerol, and lactose are present at 0.05, 0.5, and 0.2%, respectively. A modified autoinduction medium containing 0.05% of L-arabinose is used for the autoinduction of proteins in systems based on the P_{BAD} promoter (BL21AI cells) [116].

Finally, the autoinduction medium allows cultures to reach high cell densities and generally produces a greater proportion of soluble target protein than IPTG-induced expression [116]. Autoinduction reagents are commercially available (Overnight Express™, Novagen). A recent study demonstrated dosing the LacI repressor affects carbon consumption patterns, thus dramatically influencing protein expression. It was observed that, when using a system that provided high amounts of LacI (e.g., T7lac or T5lac), the order of consumption shifted from glucose/lactose/glycerol to glucose/glycerol/lactose, thus delaying the expression of the target protein [117]. Also, when using a system such as T5lacI^q, which produces even more LacI, the effect was so dramatic that culture growth stopped before lactose could be consumed. The oxygenation rate also affects consumption preferences; in cases where O_2 was limiting, lactose was consumed at lower cell densities [117].

6 HTS for expression of RPs

In the late 1970s and early 1980s, the components that made modern HTS possible in the laboratory came together. The explosive growth of HTS led to a great abundance of automation technology, ranging from simple, small, and affordable liquid-handling workstations to very large factory-style integrated systems. These fully automated systems are the most valuable tools available for HTS.

Over the last few years, a number of HTS technologies have been developed to expedite the production of RP for therapeutic studies. These include the use of rapid cloning systems, miniaturization of cell growth conditions, and a variety of innovative automation systems for expression and purification of soluble target proteins.

Based on the idea that the probability to obtain soluble RP depends on a complex array of many variables (strains, vectors, and culture conditions), an interesting approach is to try as many variables as possible in the shortest time. The implementation of this technology has often found the optimal vector, strain, and/or culturing condition required for successfully expressing and purifying the specific target protein, as well as the refolding conditions for insoluble proteins [15, 28, 118, 119].

The evaluation of hundreds of conditions can be achieved automatically by the use of robotic platforms or by manually using multichannel pipettors. The first step in the HTS dedicated to the production of soluble RP is to clone the target gene in different expression vectors, as well as to evaluate, in some cases, several truncated forms or mutant variants of the gene of interest [13, 28]. As mentioned in previous section, several methodologies have been developed to aid in the cloning of several genes in multiple vectors. The RF cloning method described by Unger and colleagues [101] is an optimal choice for this step and in our hands has proven to be a cost-effective and very efficient methodology (unpublished results). After cloning, the constructs need to be introduced into an expression host. *E. coli* is a robust organism and can be cultivated in 24- (2 mL per well) and 96-well plates (1 mL per well) covered with AirPore tape (Quiagen), thus making the expression setup very simple. The use of rich media, such as terrific broth or autoinduction media [116], to ensure maximum biomass is preferred in the HTS analysis. Generally, these media support optical densities (OD_{600}) of 5–10 units compared with 2–3 units for LB media. Referring to the optimal temperature used in these approaches, a low temperature (15–25°C) is highly recommended [16, 27, 28, 57, 120, 121].

After protein expression, cells could be harvested by centrifugation (2.500 *g* for 15 min) and resuspended in lysis buffer. The lysis buffer should contain protease inhibitors (complete EDTAfree, Roche) and high ionic strength (0.3–0.5M NaCl).

Also, if the lysates are going to be purified by IMAC, a low concentration of imidazol (20–40 mM) should be added to the lysis buffer to diminish nonspecific binding of host proteins to the resin [16].

To facilitate the automation and downstream analysis of protein expression, the centrifugation process could be skipped and bacteria could be directly lysed in the growth media by the addition of commercially available chemical reagents, such as PopCulture™ (Novagen) and FastBreak™ (Promega) [58, 122]. As an alternative, cell lysis could be also achieved by the addition of lysozyme (1 mg/mL) and freeze–thaw cycles combined with sonication. Sonication devices adapted for robotic platforms of the 96-well plate format are available (Misonix) [28].

Is important to highlight that cell lysis is a critical step when working with small cultures and requires the use of specialized equipment, chemical reagents, or freeze–thaw cycles that increase the cost or make the automation process more difficult. In this regard, novel strategies have been developed that are based on the intracellular expression of lytic genes [123]. For example, the expression of the lysis gene cassette *SRRz* from bacteriophage λ under the control of the heat-inducible promoter p_R (induction by raising temperature to 42°C) [124], or UV-inducible promoters, such as *recA* and *umuDC* (induction by UV irradiation for 8 min) [125], for cellular lysis were evaluated in the 96-well format. The reagent free, in situ, and cost-effective characteristics of this approach make this strategy a promising tool for HTS in the future [122].

Once cells are lysed, the supernatant can be clarified directly in the deep-well plate by centrifugation (3.000 *g* for 1 h), or can be loaded into a 96-well filter plate and the soluble fraction obtained by vacuum-driven filtration [126]. In this step, the supernatants can be directly evaluated for soluble-protein production [120] or the clarified supernatant can be loaded into a 96-well plate containing charged nickel resin for purification of his-tagged proteins (His MultiTrap, GE; Ni-NTA HisSorb, QUIAGEN) [28, 55, 56]. Many of the 96-well IMAC plates also support the purification of unclarified lysates, but final results to evaluate RP with expression problems and/or low solubility could be compromised. A 96-well plate into which the culture is directly loaded and allows simultaneous cell disruption, protein binding, and purification is also commercially available (His-Select iLAP, Sigma) [57]. Finally, agarose magnetic beads (MagneHis™ Ni-Particles, Promega; Ni-NTAMagnetic Agarose Beads, QIAGEN) are available as another choice for HTS [56–58, 121]. In this case, binding, washing, and elution steps are done by the use of a magnet (MagnaBot 96, Promega).

An alternative to IMAC purification in 96-well plates is to use Strep-tagII/Strep-Tactin™ purification. Sepharose resins coupled with Strep-Tactin in the 96-well plate format are commercially available (Strep-well HT 50, IBA), making this a suitable method for the high-throughput purification of Strep-tag proteins.

Once protein is eluted from the purification plate, the last automated action could be the functional evaluation of the target protein (if it is possible) as well as evaluation by SDS-PAGE or dot blot (if specific antibodies against the target protein or fused tag are available) [60]. A useful system to evaluate soluble expression of the target protein in HTS on a robotic platform is the E-PAGE™96 system (Invitrogen), which allows the evaluation of 96 different conditions in a short time [58, 122].

Overall, works of structural genomics centers dedicated to find the optimal conditions to obtain soluble RPs through HTS propose, as initial rules, the use of the *E. coli* T7 expression system (BL21-DE3 derivatives of *E. coli* strains and pET vectors, Novagen) and their posterior purifications through 96-well IMAC plates [16, 27, 28]. However, it is important to mention that, when working on a small scale, parameters such as temperature, culture conditions, and aeration, do not always scale well and some proteins may not be well expressed on a large scale and vice versa [16, 127]. Also, some soluble hits can result in soluble, but aggregated and/or nonfunctional proteins, showing the importance of performing biophysical characterization of the RP after protein expression (analytical gel filtration, static/dynamic light scattering, MALDI-TOF mass spectrometry) [13, 15, 28].

6.1 HTS on a robotic platform committed to finding optimal expression conditions for an insoluble target protein

To solve the problem of RPs, the typical expression conditions of which (BL21 *E. coli* strains, pET vectors, temperature expression of 37°C, and induction culture to optical densities of 0.6) did not result in expression in soluble form, we developed an HTS protocol in an automated system (liquid-handling workstations Genesis, Tecan). This protocol has been created to focus on the three steps that could have an impact on improving the level of soluble expression for a specific RP: (i) HTS of different constructs; (ii) HTS of different expression strains; and (iii) HTS of different culture conditions. A brief description of this approach is provided in the fol-

lowing section and more details are given in the Supporting Information.

6.1.1 HTS of different constructs

One of the standard procedures when setting out to express RPs is to screen a series of constructs to identify the optimal vector able to produce enough soluble protein. This may include the expression of a full-length molecule, the mutated target protein, as well as specific domains of RP or a chimera-fused protein. A series of fusion partners may also be investigated for their effects on driving enhanced expression or their capacity to capture and purify the target protein quickly with minimal impurities. By using traditional cloning methodologies, generating the many possible combinations and their analysis in different expression systems would be so labor intensive and time consuming that a parallel strategy of expression screening would be impractical. Thus, using the RF cloning method, we are able to obtain 9 different constructs of a target insoluble protein in 24 h. This approach involves the expression of the RP with a strong promoter, such as T7 with N- and C-terminal his-tags, weaker promoters, such as T5 with N- and C-terminal his-tags, as well as a tightly regulated promoter, such as P_{BAD} with an N-terminal his-tag. In addition to these five constructs, we prepared four other constructs with fusion tags involving GST, Mal-E, Nus-A, and Trx proteins.

6.1.2 HTS of different expression strains

Many *E. coli* strains optimized for protein expression purposes are commercially available from suppliers such as Invitrogen, Novagen, and Stratagene. These strains are sold in 8-well strips and 96-well plate formats, allowing convenient transfer of protocols to HTS formats using liquid-handling workstations. In our case, we combined these 9 constructs referred to above with 6 different *E. coli* strains, achieving a total of 54 different expression conditions of the target protein.

6.1.3 HTS of different culture conditions

As described in previous section, the culture conditions constitute another important variable that should be taken into account to improve the quantity of soluble target protein. In this context, we optimized the HTS robotic system for testing all of these 54 variables at 2 different temperatures selected as required (if the protein target is insoluble at the typical 37°C, different temperatures, such as 25 and 16°C, could be evaluated). Finally, the automated system was also optimized to obtain $O.D._{600}$ values measured every 60 min and consequently proceeded to IPTG induction at 2 different growth states of the bacterial culture (e.g., $O.D._{600} \approx 0.6$ and 1.0). If the target protein is involved in a deleterious way for the host, this strategy can often be successful to obtain some quantities of the desired protein.

These two last automated steps (construct generation, strain transformation, and bacterial growth/IPTG induction) are the first part of this HTS protocol developed in our group. Thus, 216 different variables of target protein expression can be evaluated. This first part of the HTS method is developed with minimal human intervention and is successfully achieved in 24 h.

The second part of this HTS approach involves evaluation of the soluble state of the target protein. For this, cellular lysates of 216 conditions are produced in the same 96-well plates, allowing the complete lysis of *E. coli* strains and eventual filtration directly from the culture medium. After filtration, soluble and insoluble fractions are collected and migrated in SDS-PAGE to perform western blot analysis. The analysis of these results allows the identification of the condition/s in which the target protein is expressed in a soluble form and subsequently the performance of large-scale production of the RP. This second part of HTS protocol is carried out on the robotic platform and completed in 6 h.

The described protocol allows the evaluation of 216 expression conditions (different constructs, different expression strains, and different culture conditions) in 4 days, for a protein that, so far, could not be expressed in soluble form and enough quantity. Presently, this protocol is being implemented in our laboratory in collaboration with the Structural Biochemistry Unit of the Pasteur Institute in France. The graphical pathway of this pipeline is shown in Fig. 1 and the protocol is detailed in the Supporting Information.

Despite the fact that the robotic platform allows to evaluate many variables in a short time, this approach could also be implemented on a minor scale in laboratories without robotic technology.

Taking into account these, the laboratories could test different expression variables with minimal equipment investment. For example, with the use of multichannel pipettors, 96- or 24-deep-well plates, as well as different vectors and *E. coli* strains, similar experiments could be done manually. While the success can be greater when more conditions are evaluated, this manual approach can be improved by bioinformatics studies on the target protein. Many of these protein analyses are successful in identifying exposed hydrophobic amino acids as well as rare codon usage and other features

that could determine the correct folding of target protein.

7 Conclusion

Determining protein function and structure are fundamental for the continuous progress of biotechnology in the post-genomic world. As mentioned before, even though important progress has been made, there is no magic formula that is able to ensure the production of soluble and functional target proteins. To select the proper host for the production of a particular protein, the existence of complex post-translational modifications (e.g., glycosylation), the heterologous origin, and thus codon biases, the presence of disulfide bonds or toxicity of the protein are the basic rules to take into account to get closer to, but not ensure, success.

HTS emerges as an innovative tool that allows the screening of hundreds of different conditions in a reasonable time. Processes such as cloning, expression/induction, cell lysis, protein purification, and protein visualization by SDS-PAGE/western blot were automated, making the evaluation of many hundreds of expression conditions in one week possible.

Although greatly enhancing throughput and the ability to study more conditions, the technology should not be used as a replacement for sensible experimental design. Despite the fact that HTS systems emerge as invaluable tools for the future of the RP field, we must highlight that, while it is not a universal solution for all RPs, it is an important support tool.

We specially thank Dr. Pedro Alzari and Dr. Ahmed Haouz for economic and scientific help in the improvement of the robotic platform dedicated to the soluble expression of recombinant proteins.

The authors have declared no conflict of interest.

8 References

[1] Cohen, S. N., Chang, A. C., Boyer, H. W., Helling, R. B., Construction of biologically functional bacterial plasmids in vitro. *Proc. Natl. Acad. Sci. USA* 1973, *70*, 3240–3244.

[2] Sorensen, H. P., Towards universal systems for recombinant gene expression. *Microb. Cell Fact.* 2010, *9*, 27.

[3] Terpe, K., Overview of bacterial expression systems for heterologous protein production: From molecular and biochemical fundamentals to commercial systems. *Appl. Microbiol. Biotechnol.* 2006, *72*, 211–222.

[4] Jana, S., Deb, J. K., Strategies for efficient production of heterologous proteins in *Escherichia coli*. *Appl. Microbiol. Biotechnol.* 2005, *67*, 289–298.

[5] Brondyk, W. H., Selecting an appropriate method for expressing a recombinant protein. *Methods Enzymol.* 2009, *463*, 131–147.

[6] Widmann, M., Christen, P., Comparison of folding rates of homologous prokaryotic and eukaryotic proteins. *J. Biol. Chem.* 2000, *275*, *18* 619–18 622.

[7] Dyson, M. R., Shadbolt, S. P., Vincent, K. J., Perera, R. L., McCafferty, J., Production of soluble mammalian proteins in *Escherichia coli*: Identification of protein features that correlate with successful expression. *BMC Biotechnol.* 2004, *4*, 32.

[8] Wagner, S., Klepsch, M. M., Schlegel, S., Appel, A. et al., Tuning *Escherichia coli* for membrane protein overexpression. *Proc. Natl. Acad. Sci. USA* 2008, *105*, *14* 371–14 376.

[9] Walls, D., Loughran, S. T., Tagging recombinant proteins to enhance solubility and aid purification. *Methods Mol. Biol. (Clifton, N.J)* 2011, *681*, 151–175.

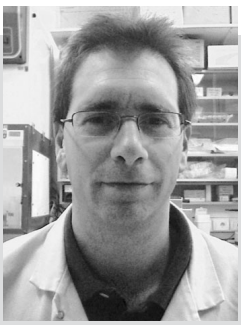
[10] Kyratsous, C. A., Silverstein, S. J., DeLong, C. R., Panagiotidis, C. A., Chaperone-fusion expression plasmid vectors for improved solubility of recombinant proteins in *Escherichia coli*. *Gene* 2009, *440*, 9–15.

[11] Martinez-Alonso, M., Garcia-Fruitos, E., Ferrer-Miralles, N., Rinas, U., Villaverde, A., Side effects of chaperone gene coexpression in recombinant protein production. *Microb. Cell Fact.* 2010, *9*, 64.

[12] Listwan, P., Terwilliger, T. C., Waldo, G. S., Automated, high-throughput platform for protein solubility screening using a split-GFP system. *J. Struct. Funct. Genomics* 2009, *10*, 47–55.

[13] Klock, H. E., Koesema, E. J., Knuth, M. W., Lesley, S. A., Combining the polymerase incomplete primer extension method for cloning and mutagenesis with microscreening to accelerate structural genomics efforts. *Proteins* 2008, *71*, 982–994.

[14] Sala, E., de Marco, A., Screening optimized protein purification protocols by coupling small-scale expression and mini-



Dr. Pablo Oppezzo received his M.Sc. in Molecular and Cellular Biology from PEDECIBA (Uruguay) in 1999. In 2004, he received his Ph.D. in Immunology from the University of Paris VI, and did post-doctoral work at the Biochemical Structural Unit at the Pasteur Institute of Paris. In 2006, he obtained a position as Principal Investigator of Recombinant Protein Unit at the Institut Pasteur de Montevideo, Uruguay. Dr. Pablo Oppezzo is now the Head of the Recombinant Protein Unit and a collaborator in the Immunology department at the Facultad de Medicina, Universidad de la República (UdelaR), Uruguay. Dr. Oppezzo's research focuses on studying the mechanisms involved in the origins of hematopoietic B-cell malignancies. In recent years, he contributed to the study of the mechanism of somatic hypermutation and class-switch recombination processes in Chronic Lymphocytic Leukemia, as well as in the development of new prognosis markers for this disease.

size exclusion chromatography. *Protein Expression Purif.* 2010, *74*, 231–235.
[15] Acton, T. B., Gunsalus, K. C., Xiao, R., Ma, L. C. et al., Robotic cloning and protein production platform of the northeast structural genomics consortium. *Methods Enzymol.* 2005, *394*, 210–243.
[16] Graslund, S., Nordlund, P., Weigelt, J., Hallberg, B. M. et al., Protein production and purification. *Nat. Methods* 2008, *5*, 135–146.
[17] Ye, H., Simultaneous determination of protein aggregation, degradation, and absolute molecular weight by size exclusion chromatography multiangle laser light scattering. *Anal. Biochem.* 2006, *356*, 76–85.
[18] Sone, M., Kishigami, S., Yoshihisa, T., Ito, K., Roles of disulfide bonds in bacterial alkaline phosphatase. *J. Biol. Chem.* 1997, *272*, 6174–6178.
[19] Nakamoto, H., Bardwell, J. C., Catalysis of disulfide bond formation and isomerization in the *Escherichia coli* periplasm. *Biochim. Biophys. Acta* 2004, *1694*, 111–119.
[20] Yoon, S. H., Kim, S. K., Kim, J. F., Secretory production of recombinant proteins in *Escherichia coli*. *Recent Pat. Biotechnol.* 2010, *4*, 23–29.
[21] Francis, D. M., Page, R., Strategies to optimize protein expression in *E. coli*. *Current protocols in protein science* 2010, 61:5.24.1–5.24.29.
[22] Mairhofer, J., Cserjan-Puschmann, M., Striedner, G., Nobauer, K. et al., Marker-free plasmids for gene therapeutic applications—lack of antibiotic resistance gene substantially improves the manufacturing process. *J. Biotechnol.* 2010, *146*, 130–137.
[23] Hagg, P., de Pohl, J. W., Abdulkarim, F., Isaksson, L. A., A host/plasmid system that is not dependent on antibiotics and antibiotic resistance genes for stable plasmid maintenance in *Escherichia coli*. *J. Biotechnol.* 2004, *111*, 17–30.
[24] Sorensen, H. P., Mortensen, K. K., Advanced genetic strategies for recombinant protein expression in *Escherichia coli*. *J. Biotechnol.* 2005, *115*, 113–128.
[25] Selzer, G., Som, T., Itoh, T., Tomizawa, J., The origin of replication of plasmid p15A and comparative studies on the nucleotide sequences around the origin of related plasmids. *Cell* 1983, *32*, 119–129.
[26] Hannig, G., Makrides, S. C., Strategies for optimizing heterologous protein expression in *Escherichia coli*. *Trends Biotechnol.* 1998, *16*, 54–60.
[27] Alzari, P. M., Berglund, H., Berrow, N. S., Blagova, E. et al., Implementation of semi-automated cloning and prokaryotic expression screening: The impact of SPINE. *Acta Crystallogr.* 2006, *62*, 1103–1113.
[28] Xiao, R., Anderson, S., Aramini, J., Belote, R. et al., The high-throughput protein sample production platform of the Northeast Structural Genomics Consortium. *J. Struct. Biol.* 2010, *172*, 21–33.
[29] Studier, F. W., Moffatt, B. A., Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J. Mol. Biol.* 1986, *189*, 113–130.
[30] Dubendorff, J. W., Studier, F. W., Controlling basal expression in an inducible T7 expression system by blocking the target T7 promoter with lac repressor. *J. Mol. Biol.* 1991, *219*, 45–59.
[31] Lost, I., Guillerez, J., Dreyfus, M., Bacteriophage T7 RNA polymerase travels far ahead of ribosomes in vivo. *J. Bacteriol.* 1992, *174*, 619–622.
[32] Studier, F. W., Rosenberg, A. H., Dunn, J. J., Dubendorff, J. W., Use of T7 RNA polymerase to direct expression of cloned genes. *Methods Enzymol.* 1990, *185*, 60–89.
[33] Grossman, T. H., Kawasaki, E. S., Punreddy, S. R., Osburne, M. S., Spontaneous cAMP-dependent derepression of gene expression in stationary phase plays a role in recombinant expression instability. *Gene* 1998, *209*, 95–103.
[34] Pan, S. H., Malcolm, B. A., Reduced background expression and improved plasmid stability with pET vectors in BL21 (DE3). *BioTechniques* 2000, *29*, 1234–1238.
[35] Brunner, M., Bujard, H., Promoter recognition and promoter strength in the *Escherichia coli* system. *EMBO J.* 1987, *6*, 3139–3144.
[36] Amann, E., Brosius, J., Ptashne, M., Vectors bearing a hybrid trp-lac promoter useful for regulated expression of cloned genes in *Escherichia coli*. *Gene* 1983, *25*, 167–178.
[37] Amann, E., Ochs, B., Abel, K. J., Tightly regulated tac promoter vectors useful for the expression of unfused and fused proteins in *Escherichia coli*. *Gene* 1988, *69*, 301–315.
[38] Brosius, J., Erfle, M., Storella, J., Spacing of the -10 and -35 regions in the tac promoter. Effect on its in vivo activity. *J. Biol. Chem.* 1985, *260*, 3539–3541.
[39] Saida, F., Uzan, M., Odaert, B., Bontems, F., Expression of highly toxic genes in *E. coli*: Special strategies and genetic tools. *Curr. Protein Pept. Sci.* 2006, *7*, 47–56.
[40] Banerjee, S., Salunkhe, S. S., Apte-Deshpande, A. D., Mandi, N. S. et al., Over-expression of proteins using a modified pBAD24 vector in *E. coli* expression system. *Biotechnol. Lett.* 2009, *31*, 1031–1036.
[41] Guzman, L. M., Belin, D., Carson, M. J., Beckwith, J., Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *J. Bacteriol.* 1995, *177*, 4121–4130.
[42] Goulding, C. W., Perry, L. J., Protein production in *Escherichia coli* for structural studies by X-ray crystallography. *J. Struct. Biol.* 2003, *142*, 133–143.
[43] Goldstein, J., Pollitt, N. S., Inouye, M., Major cold shock protein of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* 1990, *87*, 283–287.
[44] Vasina, J. A., Baneyx, F., Recombinant protein expression at low temperatures under the transcriptional control of the major *Escherichia coli* cold shock promoter cspA. *Appl. Environ. Microbiol.* 1996, *62*, 1444–1447.
[45] Inouye, S., Sahara, Y., Expression and purification of the calcium binding photoprotein mitrocomin using ZZ-domain as a soluble partner in *E. coli* cells. *Protein Expression Purif.* 2009.
[46] Liu, D., Schmid, R. D., Rusnak, M., Functional expression of Candida antarctica lipase B in the *Escherichia coli* cytoplasm—a screening system for a frequently used biocatalyst. *Appl. Microbiol. Biotechnol.* 2006, *72*, 1024–1032.
[47] Vasina, J. A., Peterson, M. S., Baneyx, F., Scale-up and optimization of the low-temperature inducible cspA promoter system. *Biotechnol. Prog.* 1998, *14*, 714–721.
[48] Qing, G., Ma, L. C., Khorchid, A., Swapna, G. V. et al., Cold-shock induced high-yield protein production in *Escherichia coli*. *Nat. Biotechnol.* 2004, *22*, 877–882.
[49] Villaverde, A., Benito, A., Viaplana, E., Cubarsi, R., Fine regulation of cI857-controlled gene expression in continuous culture of recombinant *Escherichia coli* by temperature. *Appl. Environ. Microbiol.* 1993, *59*, 3485–3487.
[50] Valdez-Cruz, N. A., Caspeta, L., Perez, N. O., Ramirez, O. T., Trujillo-Roldan, M. A., Production of recombinant proteins in *E. coli* by the heat inducible expression system based on the phage lambda pL and/or pR promoters. *Microb. Cell Fact.* 2010, *9*, 18.
[51] Menart, V., Jevsevar, S., Vilar, M., Trobis, A., Pavko, A., Constitutive versus thermoinducible expression of heterolo-

gous proteins in *Escherichia coli* based on strong PR,PL promoters from phage lambda. *Biotechnol. Bioeng.* 2003, *83*, 181–190.

[52] Porath, J., Immobilized metal ion affinity chromatography. *Protein Expression Purif.* 1992, *3*, 263–281.

[53] Li, M., Su, Z. G., Janson, J. C., In vitro protein refolding by chromatographic procedures. *Protein Expression Purif.* 2004, *33*, 1–10.

[54] Bonetta, L., Protein purification: Fast forward. *Nature* 2006, *439*, 1017–1021.

[55] Murphy, M. B., Doyle, S. A., High-throughput purification of hexahistidine-tagged proteins expressed in *E. coli*. *Methods Mol. Biol. (Clifton, N.J)* 2005, *310*, 123–130.

[56] Schafer, F., Romer, U., Emmerlich, M., Blumer, J. et al., Automated high-throughput purification of 6xHis-tagged proteins. *J. Biomol. Tech.* 2002, *13*, 131–142.

[57] Peleg, Y., Unger, T., Application of high-throughput methodologies to the expression of recombinant proteins in *E. coli*. *Methods Mol. Biol. (Clifton, N.J)* 2008, *426*, 197–208.

[58] Lin, C. T., Moore, P. A., Kery, V., Automated 96-well purification of hexahistidine-tagged recombinant proteins on MagneHis Ni(2)+-particles. *Methods Mol. Biol. (Clifton, N.J)* 2009, *498*, 129–141.

[59] Steen, J., Uhlen, M., Hober, S., Ottosson, J., High-throughput protein purification using an automated set-up for high-yield affinity chromatography. *Protein Expression Purif.* 2006, *46*, 173–178.

[60] Vincentelli, R., Canaan, S., Offant, J., Cambillau, C., Bignon, C., Automated expression and solubility screening of His-tagged proteins in 96-well format. *Anal. Biochem.* 2005, *346*, 77–84.

[61] Magnusdottir, A., Johansson, I., Dahlgren, L. G., Nordlund, P., Berglund, H., Enabling IMAC purification of low abundance recombinant proteins from *E. coli* lysates. *Nat. Methods* 2009, *6*, 477–478.

[62] Schmidt, T. G., Skerra, A., The Strep-tag system for one-step purification and high-affinity detection or capturing of proteins. *Nat. Protoc.* 2007, *2*, 1528–1535.

[63] Lichty, J. J., Malecki, J. L., Agnew, H. D., Michelson-Horowitz, D. J., Tan, S., Comparison of affinity tags for protein purification. *Protein Expression Purif.* 2005, *41*, 98–105.

[64] Stofko-Hahn, R. E., Carr, D. W., Scott, J. D., A single step purification for recombinant proteins. Characterization of a microtubule associated protein (MAP 2) fragment which associates with the type II cAMP-dependent protein kinase. *FEBS Lett.* 1992, *302*, 274–278.

[65] Raines, R. T., McCormick, M., Van Oosbree, T. R., Mierendorf, R. C., The S.Tag fusion system for protein purification. *Methods Enzymol.* 2000, *326*, 362–376.

[66] Ikeda, T., Ninomiya, K., Hirota, R., Kuroda, A., Single-step affinity purification of recombinant proteins using the silica-binding Si-tag as a fusion partner. *Protein Expression Purif.* 2010, *71*, 91–95.

[67] Harper, S., Speicher, D. W., Purification of proteins fused to glutathione S-transferase. *Methods Mol. Biol. (Clifton, N.J)* 2011, *681*, 259–280.

[68] Kaplan, W., Husler, P., Klump, H., Erhardt, J. et al., Conformational stability of pGEX-expressed Schistosoma japonicum glutathione S-transferase: A detoxification enzyme and fusion-protein affinity tag. *Protein Sci.* 1997, *6*, 399–406.

[69] Hammarstrom, M., Hellgren, N., van Den Berg, S., Berglund, H., Hard, T., Rapid screening for improved solubility of small human proteins produced as fusion proteins in *Escherichia coli*. *Protein Sci.* 2002, *11*, 313–321.

[70] Smith, D. B., Johnson, K. S., Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione S-transferase. *Gene* 1988, *67*, 31–40.

[71] Pattenden, L. K., Thomas, W. G., Amylose affinity chromatography of maltose-binding protein: Purification by both native and novel matrix-assisted dialysis refolding methods. *Methods Mol. Biol. (Clifton, N.J)* 2008, *421*, 169–189.

[72] Zhu, S., Yang, G., Yang, X., Zhao, Y. et al., Soluble expression in *Escherichia coli* of active human cyclic nucleotide phosphodiesterase isoform 4B2 in fusion with maltose-binding protein. *Bioscience, biotechnology, and biochemistry* 2009, *73*, 968–970.

[73] Cho, H. J., Lee, Y., Chang, R. S., Hahm, M. S. et al., Maltose binding protein facilitates high-level expression and functional purification of the chemokines RANTES and SDF-1alpha from *Escherichia coli*. *Protein Expression Purif.* 2008, *60*, 37–45.

[74] Kapust, R. B., Waugh, D. S., *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci* 1999, *8*, 1668–1674.

[75] Lunn, C. A., Kathju, S., Wallace, B. J., Kushner, S. R., Pigiet, V., Amplification and purification of plasmid-encoded thioredoxin from *Escherichia coli* K12. *J. Biol. Chem.* 1984, *259*, 10 469–10 474.

[76] Kim, S., Lee, S. B., Soluble expression of archaeal proteins in *Escherichia coli* by using fusion-partners. *Protein Expression Purif.* 2008, *62*, 116–119.

[77] Yanga, Y., Tiana, Z., Teng, D., Zhang, J. et al., High-level production of a candidacidal peptide lactoferrampin in *Escherichia coli* by fusion expression. *J. Biotechnol.* 2009, *139*, 326–331.

[78] LaVallie, E. R., DiBlasio, E. A., Kovacic, S., Grant, K. L. et al., A thioredoxin gene fusion expression system that circumvents inclusion body formation in the *E. coli* cytoplasm. *Bio/Technology* 1993, *11*, 187–193.

[79] Marblestone, J. G., Edavettal, S. C., Lim, Y., Lim, P. et al., Comparison of SUMO fusion technology with traditional gene fusion systems: Enhanced expression and solubility with SUMO. *Protein Sci.* 2006, *15*, 182–189.

[80] Malakhov, M. P., Mattern, M. R., Malakhova, O. A., Drinker, M. et al., SUMO fusions and SUMO-specific protease for efficient expression and purification of proteins. *J. Struct. Funct. Genomics* 2004, *5*, 75–86.

[81] Gusarov, I., Nudler, E., Control of intrinsic transcription termination by N and NusA: The basic mechanisms. *Cell* 2001, *107*, 437–449.

[82] De Marco, V., Stier, G., Blandin, S., de Marco, A., The solubility and stability of recombinant proteins are increased by their fusion to NusA. *Biochem. Biophys. Res. Commun.* 2004, *322*, 766–771.

[83] Davis, G. D., Elisee, C., Newham, D. M., Harrison, R. G., New fusion protein systems designed to give soluble expression in *Escherichia coli*. *Biotechnol. Bioeng.* 1999, *65*, 382–388.

[84] Kim, T. W., Chung, B. H., Chang, Y. K., Production of soluble human interleukin-6 in cytoplasm by fed-batch culture of recombinant *E. coli*. *Biotechnol. Prog.* 2005, *21*, 524–531.

[85] Esposito, D., Chatterjee, D. K., Enhancement of soluble protein expression through the use of fusion tags. *Curr. Opin. Biotechnol.* 2006, *17*, 353–358.

[86] Phan, J., Zdanov, A., Evdokimov, A. G., Tropea, J. E. et al., Structural basis for the substrate specificity of tobacco etch virus protease. *J. Biol. Chem.* 2002, *277*, 50 564–50 572.

[87] Kerrigan, J. J., Xie, Q., Ames, R. S., Lu, Q., Production of protein complexes via co-expression. *Protein Expression Purif.* 2010, *75*, 1–14.
[88] Romier, C., Ben Jelloul, M., Albeck, S., Buchwald, G. et al., Co-expression of protein complexes in prokaryotic and eukaryotic hosts: Experimental procedures, database tracking and case studies. *Acta Crystallogr.* 2006, *62*, 1232–1242.
[89] Li, C., Schwabe, J. W., Banayo, E., Evans, R. M., Coexpression of nuclear receptor partners increases their solubility and biological activities. *Proc. Natl. Acad. Sci. USA* 1997, *94*, 2278–2283.
[90] Georgiou, G., Valax, P., Expression of correctly folded proteins in *Escherichia coli*. *Curr. Opin. Biotechnol.* 1996, *7*, 190–197.
[91] Nishihara, K., Kanemori, M., Kitagawa, M., Yanagi, H., Yura, T., Chaperone coexpression plasmids: Differential and synergistic roles of DnaK-DnaJ-GrpE and GroEL-GroES in assisting folding of an allergen of Japanese cedar pollen, Cryj2, in *Escherichia coli*. *Appl. Environ. Microbiol.* 1998, *64*, 1694–1699.
[92] de Marco, A., Deuerling, E., Mogk, A., Tomoyasu, T., Bukau, B., Chaperone-based procedure to increase yields of soluble recombinant proteins produced in *E. coli*. *BMC Biotechnol.* 2007, *7*, 32.
[93] Esposito, D., Garvey, L. A., Chakiath, C. S., Gateway cloning for protein expression. *Methods Mol. Biol. (Clifton, N.J)* 2009, *498*, 31–54.
[94] Zhu, B., Cai, G., Hall, E. O., Freeman, G. J., In-fusion assembly: Seamless engineering of multidomain fusion proteins, modular vectors, and mutations. *BioTechniques* 2007, *43*, 354–359.
[95] Berrow, N. S., Alderton, D., Sainsbury, S., Nettleship, J. et al., A versatile ligation-independent cloning method suitable for high-throughput expression screening applications. *Nucleic Acids Res.* 2007, *35*, e45.
[96] Aslanidis, C., de Jong, P. J., Ligation-independent cloning of PCR products (LIC-PCR). *Nucleic Acids Res.* 1990, *18*, 6069–6074.
[97] Aslanidis, C., de Jong, P. J., Schmitz, G., Minimal length requirement of the single-stranded tails for ligation-independent cloning (LIC) of PCR products. *PCR Methods Appl.* 1994, *4*, 172–177.
[98] Curiel, J. A., de Las Rivas, B., Mancheno, J. M., Munoz, R., The pURI family of expression vectors: A versatile set of ligation independent cloning plasmids for producing recombinant His-fusion proteins. *Protein Expression Purif.* 2011, *76*, 44–53.
[99] Dan, H., Balachandran, A., Lin, M., A pair of ligation-independent *Escherichia coli* expression vectors for rapid addition of a polyhistidine affinity tag to the N- or C-termini of recombinant proteins. *J. Biomol. Tech.* 2009, *20*, 241–248.
[100] Eschenfeldt, W. H., Lucy, S., Millard, C. S., Joachimiak, A., Mark, I. D., A family of LIC vectors for high-throughput cloning and purification of proteins. *Methods Mol. Biol. (Clifton, N.J)* 2009, *498*, 105–115.
[101] Unger, T., Jacobovitch, Y., Dantes, A., Bernheim, R., Peleg, Y., Applications of the Restriction Free (RF) cloning procedure for molecular manipulations and protein expression. *J. Struct. Biol.* 2010, *172*, 34–44.
[102] Phillips, T. A., VanBogelen, R. A., Neidhardt, F. C., lon gene product of *Escherichia coli* is a heat-shock protein. *J. Bacteriol.* 1984, *159*, 283–287.
[103] Grodberg, J., Dunn, J. J., ompT encodes the *Escherichia coli* outer membrane protease that cleaves T7 RNA polymerase during purification. *J. Bacteriol.* 1988, *170*, 1245–1253.
[104] Studier, F. W., Use of bacteriophage T7 lysozyme to improve an inducible T7 expression system. *J. Mol. Biol.* 1991, *219*, 37–44.
[105] Salinas, G., Pellizza, L., Margenat, M., Fló, M., Fernandez, C., Tuned *Escherichia coli* as a host for the expression of disulfide-rich proteins. *Biotechnol. J.* 2011, DOI: 10.1002/biot.20100033
[106] de Marco, A., Strategies for successful recombinant expression of disulfide bond-dependent proteins in *Escherichia coli*. *Microb. Cell Fact.* 2009, *8*, 26.
[107] Derman, A. I., Prinz, W. A., Belin, D., Beckwith, J., Mutations that allow disulfide bond formation in the cytoplasm of *Escherichia coli*. *Science* 1993, *262*, 1744–1747.
[108] Prinz, W. A., Aslund, F., Holmgren, A., Beckwith, J., The role of the thioredoxin and glutaredoxin pathways in reducing protein disulfide bonds in the *Escherichia coli* cytoplasm. *J. Biol. Chem.* 1997, *272*, 15 661–15 667.
[109] Stewart, E. J., Aslund, F., Beckwith, J., Disulfide bond formation in the *Escherichia coli* cytoplasm: An in vivo role reversal for the thioredoxins. *EMBO J.* 1998, *17*, 5543–5550.
[110] Hatahet, F., Nguyen, V. D., Salo, K. E., Ruddock, L. W., Disruption of reducing pathways is not essential for efficient disulfide bond formation in the cytoplasm of *E. coli*. *Microb. Cell Fact.* 2010, *9*, 67.
[111] Miroux, B., Walker, J. E., Over-production of proteins in *Escherichia coli*: Mutant hosts that allow synthesis of some membrane proteins and globular proteins at high levels. *J. Mol. Biol.* 1996, *260*, 289–298.
[112] Gustafsson, C., Govindarajan, S., Minshull, J., Codon bias and heterologous protein expression. *Trends Biotechnol.* 2004, *22*, 346–353.
[113] Maertens, B., Spriestersbach, A., von Groll, U., Roth, U. et al., Gene optimization mechanisms: A multi-gene study reveals a high success rate of full-length human proteins expressed in *Escherichia coli*. *Protein Sci.* 2010, *19*, 1312–1326.
[114] Rosano, G. L., Ceccarelli, E. A., Rare codon content affects the solubility of recombinant proteins in a codon bias-adjusted *Escherichia coli* strain. *Microb. Cell Fact.* 2009, *8*, 41.
[115] Vera, A., Gonzalez-Montalban, N., Aris, A., Villaverde, A., The conformational quality of insoluble recombinant proteins is enhanced at low growth temperatures. *Biotechnol. Bioeng.* 2007, *96*, 1101–1106.
[116] Studier, F. W., Protein production by auto-induction in high density shaking cultures. *Protein Expression Purif.* 2005, *41*, 207–234.
[117] Blommel, P. G., Becker, K. J., Duvnjak, P., Fox, B. G., Enhanced bacterial protein expression during auto-induction obtained by alteration of lac repressor dosage and medium composition. *Biotechnol. Prog.* 2007, *23*, 585–598.
[118] Vincentelli, R., Canaan, S., Campanacci, V., Valencia, C. et al., High-throughput automated refolding screening of inclusion bodies. *Protein Sci.* 2004, *13*, 2782–2792.
[119] Eshaghi, S., Hedren, M., Nasser, M. I., Hammarberg, T. et al., An efficient strategy for high-throughput expression screening of recombinant integral membrane proteins. *Protein Sci.* 2005, *14*, 676–683.
[120] Busso, D., Stierle, M., Thierry, J. C., Moras, D., Automated recombinant protein expression screening in *Escherichia coli*. *Methods Mol. Biol. (Clifton, N.J)* 2008, *426*, 175–186.
[121] Sauder, M. J., Rutter, M. E., Bain, K., Rooney, I. et al., High throughput protein production and crystallization at NYSGXRC. *Methods Mol. Biol. (Clifton, N.J)* 2008, *426*, 561–575.

[122] Koehn, J., Hunt, I., High-Throughput Protein Production (HTPP): A review of enabling technologies to expedite protein production. *Methods Mol. Biol. (Clifton, N.J)* 2009, *498*, 1–18.

[123] Lin, Z., Cai, Z., Cell lysis methods for high-throughput screening or miniaturized assays. *Biotechnol. J.* 2009, *4*, 210–215.

[124] Cai, Z., Xu, W., Xue, R., Lin, Z., Facile, reagentless and in situ release of *Escherichia coli* intracellular enzymes by heat-inducible autolytic vector for high-throughput screening. *Protein Eng., Des. Sel.* 2008, *21*, 681–687.

[125] Li, S., Xu, L., Hua, H., Ren, C., Lin, Z., A set of UV-inducible autolytic vectors for high throughput screening. *J. Biotechnol.* 2007, *127*, 647–652.

[126] Knaust, R. K., Nordlund, P., Screening for soluble expression of recombinant proteins in a 96-well format. *Anal. Biochem.* 2001, *297*, 79–85.

[127] Kim, Y., Bigelow, L., Borovilos, M., Dementieva, I. et al., *Advances in Protein Chemistry and Structural Biology* 2008, *75*, pp. 85–105.

ANEXO II:

Overcoming the solubility problem in *E. coli*: available approaches for recombinant protein production.

Agustín Correa[1], Pablo Oppezzo[1]

[1] Recombinant Protein Unit, Institut Pasteur de Montevideo, Uruguay.

Methods in Molecular Biology: "Insoluble Proteins" book series. Chapter 1. **Accepted**

Chapter 2

Overcoming the Solubility Problem in *E. coli*: Available Approaches for Recombinant Protein Production

Agustín Correa and Pablo Oppezzo

Abstract

Despite the importance of recombinant protein production in academy and industrial fields, many issues concerning the expression of soluble and homogeneous product are still unsolved. Although several strategies were developed to overcome these obstacles, at present there is no magic bullet that can be applied for all cases. Indeed, several key expression parameters need to be evaluated for each protein. Among the different hosts for protein expression, *Escherichia coli* is by far the most widely used. In this chapter, we review many of the different tools employed to circumvent protein insolubility problems.

Key words Recombinant proteins, Protein expression, *E. coli*, High-throughput screening, Inclusion bodies, Directed evolution

1 Introduction

With the advances in genome sequencing nowadays, over 1,900 genomes are publicly available (http://www.microbesonline.org) generating massive information in this area. A typical microbial genome codes for between 1,500 and 8,000 proteins while in eukaryotic genomes is around 10,000–60,000 proteins. Despite all this information, and in contrast with nucleic acids, obtaining the target protein from the natural host in a soluble, homogeneous state and enough quantities for biochemical and structural studies is very uncommon. This makes the production of the target protein in a recombinant form the method of choice. Different expression hosts are available for recombinant expression, including bacterial, fungal, or eukaryotic host cells. Among these, the use of the enterobacterium *Escherichia coli* is the most commonly employed with approximately 60 % of all recombinant proteins in the literature and nearly 30 % of the currently approved recombinant therapeutic proteins produced on it [1, 2]. This is mainly due to the low cost, fast growth, easy handling, high yield of target protein and the extensive knowledge of the genetics of *E. coli*.

Elena García-Fruitós (ed.), *Insoluble Proteins: Methods and Protocols*, Methods in Molecular Biology, vol. 1258,
DOI 10.1007/978-1-4939-2205-5_2,

However, when working with eukaryotic proteins, it has being estimated that approximately only 30 % of the cloned genes can be expressed in a soluble form in *E. coli* where the rest of the targets are degraded, expressed as insoluble aggregates known as inclusion bodies (IBs) or undetectable in cell extracts [3]. This is especially the case for membrane proteins or those requiring posttranslational modifications for folding or function. In order to overcome these limitations, several *E. coli* strains were developed as well as vectors carrying promoters with different strengths, fusion of the gene of interest with molecular tags that can aid in the purification and/or soluble production of the target protein, or the co-expression of chaperones or biological partners that can improve protein folding and stability [4, 5]. Furthermore, with the advent of the high-throughput screening (HTS) technology, all these variables can be evaluated in a simultaneous, fast, automated, and reliable manner in order to find the combination of the parameters that enable a soluble protein production [6]. Despite all this, soluble and homogeneous expressions of the target protein are not always the case. In this regard, many efforts were done with some success in the refolding of insoluble proteins from IBs [3, 7].

As an alternative strategy, the introduction of rational or, moreover, random mutations into the gene of interest in order to obtain a variant with stabilized properties or increased soluble expression has shown to be an attractive and effective approach in the soluble expression of target proteins, thus being a suitable last resource when everything else fails [8, 9].

2 Common Problems When Expressing Recombinant Proteins in *E. coli*

One of the main reasons why eukaryotic proteins often fail to be produced as soluble proteins in *E. coli* is the requirement of posttranslational modifications for correct folding. So a first step could be a sequence-based prediction of these modifications. In this regard, the ExPASy server (http://www.expasy.org) contains numerous bioinformatic tools that can estimate with a good accuracy the presence or not of posttranslational modifications like N- or O-glycosylation sites, phosphorylation sites, and protein localization, among others [10]. All this information can give us an idea of the possible success and help us in the strategy to follow for protein expression. Other factors that can have an impact in the soluble expression of the target are the codon usage, the sequence at the translation initiation region (TIR), as well as the correct formation of disulfide bridges. A brief description of the strategies designed to obtain soluble recombinant proteins is given in the next sections, and finally, in-depth information of them will be given in the following chapters of this book.

2.1 Effects of DNA/RNA Sequence in Protein Expression

The presence of uncommon codons for *E. coli* can have a strong influence in the gene expression. Because of the heterologous nature of the target protein, the target gene may have codons that are in low abundance in this host. This can lead into growth arrest, premature translation termination, and low yield of protein production, among others [11].

In order to overcome this problem, two different approaches have been proposed: (1) the substitution of the rare codons present in the gene sequence by de novo gene synthesis or (2) the expression of the gene of interest in an *E. coli* strain that is supplemented by tRNAs that are present in low abundance. In the former case, several algorithms were developed in order to optimize the gene sequence to the host codon usage [12, 13]. More recently, a software was developed in order to not only evaluate the codon frequency but also the codon pair usage or codon context. This approach suggests that codon pair usage and codon context can be as important as the individual codon usage [14]. For the second strategy, several commercially available strains have been developed that co-express tRNAs for rare codons, like BL21 CodonPlus (Novagen) and Rosetta™ (Invitrogen). The use of such strains demonstrated to be effective for the soluble expression of several targets [15, 16].

Finally, changing the rare codons can increase the translation rate, but in some cases this can lead to protein aggregation and misfolding as it was demonstrated for several proteins expressed in *E. coli* [17]. This suggests that translation pauses can be necessary in some cases for proper folding of individual domains [18]; thus the procedure of gene optimization or the use of a codon optimized for an *E. coli* strain cannot be used as a general rule.

Also at the DNA sequence level, it has been shown that the sequence at the 5′ of the gene can have an important impact in the levels of protein expression due to the generation of secondary structures in the messenger RNA that can hamper the translation by the ribosome complex. In this regard, it was shown that sequences immediately after the start codon up to position +25 can have a profound effect in protein expression. For these reasons, there are some programs that enable the optimization of the TIRs in order to improve protein expression by defining silent mutations in the first seven codons [19]. More recently a predictive method for designing synthetic ribosome binding sites was developed where different translation initiation rates can be targeted, thus enabling the rational control and fine tuning of recombinant protein expression [20, 21].

In the same way, in bacteria, the half-life of mRNA is much shorter than in eukaryotic cells. It was shown that mutation in the gene coding for RNaseE confers increased mRNA stability [22]. A BL21 derivative strain containing such mutation is commercialized by Invitrogen under the name of BL21 Star.

2.2 Disulfide Bonds: A Common Posttranslational Modification Implies a Common Problem for Recombinant Protein Expression

Disulfide bonds correspond to a covalent linkage between two sulfur atoms from two cysteine residues. They are frequently essential for proper folding, stability, and/or function of the target protein, thus a very important feature to take into account when expressing a target gene [23]. The presence of disulfide bonds can be predicted by web-based servers in order to estimate if the target protein can require such posttranslational modification [24, 25].

Disulfide bonds are formed in oxidizing environments like the eukaryotic endoplasmic reticulum or the bacterial periplasm. The formation of disulfide bridges in the periplasmic of E. coli, requires the action of DsbC system were DsbA catalyze disulfide formation whiles DsbC catalyze the isomerization of incorrectly formed disulfide bridges. The cycle can be restarted by the actions of the membrane proteins DsbB and DsbD that recycle DsbA and DsbC, respectively [26]. The expression of recombinant proteins in the periplasm of *E. coli* has allowed the correct formation of disulfide bridges of several targets [26, 27]. Purification of proteins from the periplasm is usually easier than purification of proteins from total cell lysates, since the periplasm contains a less complex protein mixture than the cytoplasm [28]. Targeting proteins to the periplasm of *E. coli* can be achieved by the addition of an N-terminal leader peptide that, depending on its nature, can use the Sec (relatively slow, posttranslational translocation) or the SRP (fast, cotranslational translocation) pathways that transport proteins through the inner plasma membrane as unfolded precursors [29, 30]. There is another translocation system: the twin-arginine translocation pathway, named Tat pathway, that, in contrast with the aforementioned pathways, catalyzes the translocation of proteins in their folded state [31].

However, one common drawback of periplasmic expression is that the translocation machinery can be saturated, which can be toxic for the host cell and decrease the final yield of the target protein. By using a strain where the expression intensity can be precisely controlled like Lemo21(DE3) (New England Biolabs), the saturation of the translocation machinery can be avoided, and thus these negative effects are minimized [32, 33].

As an alternative to periplasmic expression, engineered *E. coli* strains that contain a more oxidizing cytoplasm were developed in order to improve disulfide bond formation in this compartment. These strains contain mutations in the genes for glutathione reductase (*gor*) and thioredoxin reductase (*trxB*) involved in the maintenance of the reduced environment in the cytoplasm and a mutation in the peroxiredoxin gene *ahpC* essential for restoring growth in these mutants [23, 34]. One strain containing such mutations and used for the expression of disulfide bridges containing proteins is Origami, commercialized by Novagen [35]. However, a common problem for using such strains is the lack of disulfide bond isomerization. In this regard, a strain containing the trxB/gor/ahpC

mutations that express the DsbC isomerase in the cytoplasm of *E. coli* was developed and commercialized by New England Biolabs known as Shuffle, allowing the soluble expression of some disulfide-containing proteins in its cytoplasm [36]. Recently, by the co-expression of the sulfhydryl oxidase from *S. cerevisiae* Erv1p and the *E. coli* disulfide isomerase DsbC, disulfide bonds were generated in proteins expressed in the *E. coli* cytoplasm with the reducing pathways intact. Moreover, for some cases it was shown that the addition of a catalyst for the formation of disulfide bonds could be more effective than the removal of the reducing pathways [37, 38]. In the same sense, after making N-terminal fusions with DsbC with 28 different small disulfide-rich proteins, it was found that the strain BL21(DE3)pLysS was much more efficient in producing soluble and oxidized folded proteins in comparison to Origami B(DE3)pLysS or Shuffle T7 Express lysY cells [39]. Interestingly, when one of the fusions was used to evaluate if the disulfide bound formation occurred in the cytoplasm of BL21(DE3)pLysS cells or during the extraction and purification steps, it was found that this process occurred ex vivo [39].

3 Boosting Protein Purification and/or Expression

3.1 The Use of Fusion Tags/ Proteins

With the advent of genetic engineering, the target gene can be easily cloned in frame with different affinity and/or solubility-enhancing tags that can be exploited to increase protein solubility and yield and facilitate protein purification or downstream processing. In this regard, we can separate the fusion tags in three main categories. In the first category, referred as affinity tags, we found short tags that can be placed as N-terminus or C-terminus of the partner and can be recognized by special matrices or molecules serving for affinity purification of the fusion protein. In the second category, extremely soluble proteins with chaperone activities or thermostable characteristics in some cases are fused in order to transfer some of these properties to the fusion partner and improve the folding and/or the final yield of the target protein. Usually these tags are expressed as N-terminal fusions and are termed solubility-enhancing tags. Finally, we also have proteins that can offer a double purpose, in one hand, can be recognized by other molecules, thus serving for purification purposes, and, in the other hand, can improve the soluble production of the target protein, thus improving the target protein purity and yield [4, 40, 41].

3.2 Affinity Tags

Among the affinity tags, the His-tag is one of the most commonly used for purification of recombinant proteins in *E. coli*. This small tag (0.84 kDa) consists of 6–10 histidines in tandem and can reversibly interact with metal ions most commonly Ni or Co immobilized in a metal chelate matrix (Ni-NTA, Qiagen; Sepharose 6,

GE; or Talon resins, Clontech) [42], thus allowing mild elution conditions like the use of a competitor such as imidazole. The His-tag has several advantages like its small size and relatively inert nature, making it compatible with most downstream applications. Because the ternary structure of the His-tag is not necessary for metal coordination, it is possible to purify the protein in denaturing conditions or even perform the refolding procedure on column [43, 44]. Also the purification scheme has been automated in small and large-scale formats and has been used widely in HTS protocols, demonstrating the versatility of this tag [45–47].

As a disadvantage, when working with low-yield expressed proteins, it was shown that increasing the culture volume does not correlate with an increase in recovery. Moreover, there is a decrease in recovery because of the presence of small chelators mainly associated with the periplasm of *E. coli* that can decrease the binding capacity of the purification resin [48]. This can be improved by removing the periplasmic material before cell lysis [48]. Also it was shown that several histidine-rich *E. coli* proteins can bind to the column (like ArnA, SlyD, and GlmS), especially when working with low-expressing protein targets [49]. This reduces the purity of the target protein, consequently requiring the addition of more purification steps, thus reducing the final yield.

In this regard, some *E. coli* strains that are mutants in some of these proteins have been developed in order to overcome this issue [50, 51], and one is commercially available as NiCo21 (New England Biolabs).

Another strategy is the use of an alternative affinity tag such as Strep-tag II. This is also a small tag consisting of eight residues (WSHPQFEK) and can be specifically recognized by an engineering version of streptavidin (Strep-Tactin) [52]. Elution can be done as for the case of His-tag using mild conditions by competition with D-desthiobiotin for the Strep-tag II. Despite the binding capacity of the Strep-Tactin containing media can be lower when comparing to Sepharose 6 resins for His-tagged proteins, for example, its greater specificity makes it a good option when working with proteins that are expressed in very low quantities [52, 53]. Purification schemes for the Strep-tag II include prepacked columns as well as 96× well plates (www.iba-lifesciences.com). A variation of the Strep-tag II named Twin-Strep-tag® was recently developed and exhibited higher stability and affinity for the interaction with the Strep-Tactin. This tag consists of two *Strep*-tag®II-binding sequences connected by a short linker and showed to be more suitable for purification of diluted samples [54].

3.3 Solubility-Enhancing Tags

A common strategy to overcome the solubility problem is to fuse the target protein with a very stable and soluble one that can drive the resulting expression. It was shown for many proteins that were not soluble when expressed alone, that when expressed

as a fusion with other protein can be produced in a soluble and homogeneous state. Moreover, after cleavage and removal of the fusion partner, the target remained soluble demonstrating the utility of this approach [6, 39, 55, 56].

Among the commonly used solubility-enhancing fusion proteins, we can find the maltose-binding protein (MBP), glutathione S-transferase (GST), thioredoxin A (TrxA), disulfide isomerase C (DsbC), small ubiquitin-like modifier protein (SUMO), and N-utilization substance A (NusA).

An attractive feature of MBP and GST is that they can be used also as affinity tags. MBP is a 42 kDa protein expressed in the *E. coli* periplasm and can bind strongly to amylose resins, and elution can be done with free maltose [57]. For the case of GST, it is a 26 kDa protein from *Schistosoma japonicum* that can bind to glutathione resins, and elution is achieved by the application of reduced glutathione allowing a single-step purification process [58]. Despite GST protein is widely used, it has been shown to be a poor solubility enhancer, since in many cases after cleavage of the fusion, the target protein precipitates [6, 55, 59]. However, expression can be improved for some proteins or peptides, and the purification by glutathione resins makes it still an attractive option. Vectors for the expression of GST fusions can be found in the pGEX series from GE Healthcare or pET41a-c/pET42a-c from Novagen.

MBP was fused to either N- or C-terminus, where the expression and folding of eukaryotic fusion proteins was increased in many cases [59–61]. Vectors for MBP fusion can be found in the pMAL series from New England Biolabs or pIVEX from Roche. Also if the natural signal peptide of MBP is present ($MalE_{ss}$), expression can be directed to the periplasm of *E. coli*. This was used recently for the successful expression of disulfide-rich venom peptides [27].

TrxA is an 11.6 kDa *E. coli* thermostable (Tm: 85 °C) oxidoreductase that is expressed in very high yields. When used as a fusion tag, some of these properties can be transferred to the target protein improving its folding, solubility, and stability [62–64]. Moreover in a comparative study, all positive hits with Trx-fusions, remained still soluble after tag cleavage [6]. Expression vectors for fusion with Trx are pET32a-c from Novagen.

SUMO is a yeast protein (11.2 kDa) that when used as N-terminal fusion protein during prokaryotic expression can promote folding and soluble expression of the target protein [65–67]. Another advantage of this fusion is that it can be cleaved by a specific and efficient protease (yeast Ulp1) which recognize tertiary structure elements and a Gly-Gly-containing motif in the C-terminus of SUMO and can leave a native N-terminus on the target (except for proline) [66].

The *E. coli* disulfide isomerase DsbC (25 kDa) has isomerase and chaperonin activities [34] and has been successfully used as a fusion partner for the soluble expression of disulfide-containing

targets as mentioned earlier [39, 68]. The pET40 (Novagen) expression vector allows fusion with DsbC.

Finally, the transcription elongation and anti-termination factor of *E. coli* NusA (55 kDa) have also demonstrated to be useful for enhancing soluble protein expression [69]. In a comparative study after using several aggregation-prone target proteins, it was shown that the solubility-enhancing properties of NusA were comparable and similar to the well-studied MBP validating its utility [70]. Fusion with NusA can be achieved with the pET43.1a-c and pET44a-c vector series from Novagen.

Because TrxA, SUMO, DsbC, and NusA do not facilitate purification on their own, they are used in conjunction with small affinity tags like the aforementioned His-tag or Strep-tag II to enable protein purification. It is important to underline that despite some trends in fusion proteins were found in several studies, there is no rule for which is the best suited for the protein of interest, so it is better to test several different fusions in order to find the best option.

3.4 Tag Removal

Once the protein is expressed, in most of the cases, it is necessary to remove the fusion tag. This can be achieved by incorporating an aminoacidic sequence between the fusion tag and the protein of interest that can be recognized by a specific protease. Several proteases appear as possible options for tag removal like enterokinase (DDDDK′X), factor Xa (IE/DGR′X, where X can be any residue except for R or P), thrombin (LVPR′GS), PreScission™ protease (GE Healthcare, LEVLFQ′GP), and tobacco etch virus (TEV) protease (ENLYFQ′G), among others [41]. Between these, TEV protease is a very specific protease with several advantages like that it can be produced in the lab with high yields in *E. coli* [71], and cleavage can be done at 4 °C. Moreover, despite reducing conditions are optimal for cleavage (usually 1 mM DTT), if avoided, cleavage can still occur [27] which is preferable for disulfide-containing proteins. Also, it was demonstrated that the last glycine residue from the cleavage recognition site can be substituted by all residues except for proline, but at the expense of cleavage efficiency, allowing the release of a target protein with a native N-terminus [72].

Finally, fusion proteins were not only used for expression/purification purposes, but they were also used to obtain the crystallographic structures of several targets. This last brings the additional advantage that if the structure of the fusion is known, it can also help in the structure determination process of the target protein. Such is the case for some fusions with MBP, GST, Trx, and GFP, among others [73–76].

3.5 Cloning Methods

In order to succeed in the soluble expression of a “difficult” target protein, a recommended strategy is to test different fusion proteins, which requires the cloning of the gene of interest in several vectors. Doing this by restriction-based methods can be a complicated task,

principally when different restriction sites are present and even further when working with several targets at the same time. Nowadays some methodologies were developed as alternatives to the restriction-based cloning to facilitate the easy transfer of a DNA fragment into several vectors. Commercial kits like Gateway (Invitrogen) [77] and In-FusionTM (Clontech) [78] are efficient recombination-based cloning methods. For the case of Gateway, a suite of vectors for the easy transfer of the same DNA fragment between vectors is available. More recently, a cloning method based only in PCR reactions was developed and initially termed as RF cloning (RF, restriction free) [79]. In this method, the DNA is amplified with primers that contain complementary sequences with the site of insertion in the destination plasmid. So after the first PCR, the generated megaprimer is used in a second PCR to amplify the whole plasmid, inserting in this reaction the gene of interest in the desired position. The advantage is that insertion can be done at any position in the destination vector avoiding extra sequences to be added to the gene of interest, and if several vectors contain the same insertion sequence, the same megaprimer can be used in all of them, facilitating the cloning stage and allowing an automated HTS cloning approach [4, 79]. So by using a vector containing a fusion protein, just by inserting in the same position of the fusion other genes (like MBP, GST, SUMO, etc.), one can easily make its own suite of expression vectors where the site of insertion for the target gene is conserved along all vectors [80]. Recently, an improved protocol for RF cloning termed Transfer-PCR was developed where the generation of the megaprimer and subsequent integration of the PCR product into the destination vector occur in a single PCR reaction [81]. A web-based tool was developed for the correct design of the primers for RF cloning and is freely available (http://www.rf-cloning.org) [82]. The use of this kind of tools for molecular cloning is very useful for the generation of the genetic constructs necessary for finding a condition for soluble expression.

3.6 Expression Conditions

At the culture level, several parameters like induction temperature and medium composition can have an important effect in soluble protein yields. It was shown that lower temperature during induction (16–25 °C) can increase the final yield of soluble protein. It was assumed that a slower translation rate could favor the correct folding of the protein [83]. However, the lower temperature can also decrease the final biomass, so if the protein is well expressed, this can hamper the final yield [6]. In general, it is necessary to evaluate different temperatures to find the optimal condition. At the medium level, several media have been used for protein expression: Luria Broth (LB), 2xYT, Terrific Broth (TB), Super Broth (SB), autoinduction medium, and others. Among these media, the autoinduction medium, developed by Studier [84], has been used with success for protein expression screening in a wide range of

scales because it produces a high level of biomass. Thus, there is no need to monitor the growth; induction of cultures in well plates occurs at a comparable growth phase, which is preferable in HTS experiments; and there is a tighter control of protein induction improving expression of toxic proteins [6, 84]. A disadvantage of this medium is that it is adversely affected by aeration level. This can be reduced by decreasing the level of lacI repressor provided by the expression vector [85]. Recently, it was demonstrated that the oxygen sensitivity of expression in autoinduction medium can be practically obviated. This was achieved by using a glucose fed-batch-based autoinduction medium where the glycerol carbon source was substituted with the EnBase system [86]. This system is based on a soluble polysaccharide component within the medium and slow release of the glucose units from the polymer chain by an added specific biocatalyst [85, 87]. This kind of rich media allows an increase in the biomass production, so expression conditions can be evaluated in a reduced format like a 24× deep wells, enhancing the sensibility of automated HTS screenings for soluble protein production [4, 85, 88]. Also and as it was mentioned along the text, several strains should be used in order to find the proper condition; thus a combination of temperature and strain should be included in the screening. These in conjunction with the use of different constructs (i.e., fusion tags) make a considerable number of conditions to evaluate. In this regard, the HTS methods have had a pivotal role in making this kind of approaches possible [4, 6, 88, 89].

4 Inclusion Bodies' Renaturation

Frequently proteins accumulate, as insoluble aggregates in the cytoplasm or periplasm of *E. coli* known as inclusion bodies (IBs). As dramatically as it seems, this is not always a negative issue. Some advantages of expressing the protein as IBs are the high yield of its expression and the homogeneity in composition where in some cases the recombinant target can account for more than 90 % of the proteins in that fraction, facilitating the purification of the target after renaturation [90]. Renaturation conditions involve the evaluation of several parameters like pH, ionic strength, temperature, and addition of low molecular weight compounds, among others. In this regard, several approaches in a 96× well format have been developed to facilitate the optimization of the refolding conditions, and automated HTS protocols for protein refolding were proposed [7, 91]. Apart from the mentioned parameters, these can be combined with several methods to perform the refolding process like dilution, dialysis, or in-column refolding methods [7, 43, 92].

An attractive and counter-intuitive strategy is to introduce a tag that reduces the solubility of the fusion protein and can direct the expressed protein into insoluble IBs. This is particularly useful if the target protein is toxic to the host when soluble and correctly folded. In this regard, a mutant variant of the N-terminal autoprotease N^{pro}, of classical swine fever virus termed EDDIE, when fused to the N-terminus of the target protein can reduce its solubility in such a way that the fusion accumulates as IBs. When changing from chaotropic to kosmotropic conditions, the protease becomes active and can perform the autocleavage of the fusion, leaving a native N-terminus in the target protein [93, 94]. The comprehension of IBs nature has dramatically changed in the last years. Often, it was assumed that IBs were made of inert aggregates composed of denatured or partially folded polypeptides rather from mature native molecules. Nevertheless, in the last decades, it was shown in several cases that IBs can be made with native and active proteins [95–98]. This opens the possibility of using them in downstream applications without the need of performing protein renaturation in applications where protein aggregation is not an impediment, thus facilitating production/purification and reducing costs [99–101].

5 Protein Characterization

Obtaining the protein in a soluble state does not assure proper folding of the target protein. A common scenario is to find that the protein is soluble but forms aggregates. This is indicative of unfolded regions. A last purification step by size-exclusion chromatography (SEC) is recommendable to not only remove some remaining impurities but also to assess the oligomeric state of the sample. Protein quality assessment, can be implemented at the analytical level, with microgram quantities of protein by coupling for example, Ni Sepharose 6 beads or His MultiTrap FF 96-well plates (GE Healthcare) with the minicolumns for analytical SEC (ASEC), Superdex™ 5/150 GL (GE Healthcare), when still evaluating different expression conditions [88, 102, 103]. By using an autosampler for ASEC, the characterization step can be completely automated requiring only 14 min for each sample [102]. Also, sometimes it is necessary to evaluate different combination of additives, like for the case of membrane proteins, a combination of different detergents and/or lipids and genetic constructs in order to find a condition that gives a soluble and homogeneous sample. This kind of screening requires the purification of microgram to milligram of protein. A very useful alternative is to make GFP covalent fusions with the target protein and performing fluorescence-detection size-exclusion chromatography (FSEC). By using this approach, it is possible to determine the soluble expression

level, oligomeric state, thermostability, and approximate molecular mass using only nanogram quantities of unpurified protein, allowing working directly with the soluble extracts [104, 105]. Recently, a similar approach was developed where instead of fusing the target protein with GFP, a special fluorescent probe that can specifically recognize the His-tag was used, thus overcoming the limitations that can be associated in some cases with GFP fusions like the presence of false positives or protein aggregation issues following fusion cleavage [106].

6 Directed Evolution for Soluble Protein Expression

Despite the evaluation of many expression and growth conditions, it is often impossible to obtain the target protein in a soluble and stable manner. Under these circumstances, instead of exploring more expression parameters, one can change the physical properties of the target by making mutations or deletions in the target sequence in order to improve the solubility/stability of the recombinant protein. When structural and functional information are available, these sequence modifications can be achieved by rationally designed site-directed mutagenesis [107, 108]. Unfortunately, for most of the interesting targets, structural information is not available so rational design is not possible. In these cases, an interesting alternative is the use of directed evolution. This approach is based on an iterative process consisting of a first step of sequence diversification followed by a second step of selection of the improved mutants. The diversification process is usually achieved by random mutagenesis (error-prone PCR, chemical mutagenesis, or a mutator *E. coli* strain) [109] and/or in vitro recombination (DNA shuffling) [110]. In the directed evolution approaches, a library of mutants generated by a random process is screened for the solubility/stability of the target protein. So, after mutation occurs, one must select those few mutants with the improvements in the desired phenotype among the millions of futile mutants generated. In this regard, one can perform the selection by analyzing the activity of a reporter protein (reporter tag) or in special cases the activity of the target protein [111].

One folding reporter tag that was used successfully for the evolution of active and soluble mutant variants is the GFP-folding reporter [112, 113]. In this system, the test protein is expressed as an N-terminal fusion with GFP. So the fluorescence of *E. coli* cells is directly related to the productive folding of the fused protein [112]. In this way, the isolation of the brightest cells in the search for the mutations that improve solubility can be done using simple colony-plating techniques or fluorescence-activated cell sorting (FACS) in a flow cytometer. Later this system was improved even further by the design of a self-complemented split GFP [114]

derived from an exceptionally well-folded variant of GFP, "superfolder GFP" [115]. In this case, the target protein is fused as an N-terminal fusion to a small GFP fragment (residues 215–230, GFP11), while the GFP detector fragment (residues 1–214, GFP1-10) is expressed separately in another vector. So if the target protein is expressed in a soluble form, the GFP11 fragment can interact with GFP1-10, leading to the development of fluorescence [114].

In a different approach, the target protein can be expressed as an N-terminal fusion with a selectable marker such as the chloramphenicol acetyltransferase (CAT; 25 kDa), thus conferring resistance to chloramphenicol. It was observed that if the fusion protein is expressed in a soluble form, the cell is resistant to higher concentrations of chloramphenicol than when it is expressed in an insoluble form [116]. By using this method, it was possible to obtain soluble variants of the membrane-associated human cytochrome P450 (1A2), confirming the usefulness of this method [117].

More recently another antibiotic was used as a selectable marker but in a split manner linking in vivo protein stability to antibiotic resistance. In this case the target protein is inserted into the TEM1-β-lactamase (resistance to β-lactam antibiotics) as part of a tripartite fusion [8]. The antibiotic-resistance gene is separated between residues 196 and 197, for the insertion of the target protein gene. Thus, when protein is expressed in a soluble and stable form, the two fragments of β-lactamase can interact and thereby confer resistance to β-lactam antibiotics [8]. This method showed a low false-positive rate and, as for the CAT, is based on a selection rather than a screening for obtaining improved mutants.

Another elegant approach is the colony filtration (CoFi) blot. This is based in the fact that IBs can be separated from soluble proteins by filtration at the colony level. So after transforming bacteria with the mutant library, colonies are transferred to a filter membrane where protein expression is induced and cells are then lysed. Soluble proteins can diffuse through the filter and bind to the nitrocellulose membrane for detection [118, 119]. An anti-His antibody can be used for the detection of His-tagged soluble variants making it an easy to adopt method. Cornvik and colleagues randomized the N-terminal region of 32 mammalian proteins, and mutants were selected for soluble expression using this methodology. By this approach, the success rate for soluble expression was increased from 34 to 68 %, showing the high potential of this methodology [118].

Just as in the HTS, usually many different expression conditions for the same protein are evaluated; in the directed evolution approach, a library of mutants generated by a random process is screened for the solubilization/stabilization of the target protein. The key issues in this strategy are the diversity of the library and the selection/isolation method employed for finding the mutant with the improved characteristics.

7 Conclusions and Future Perspectives

Although a lot of progress has been made in recombinant protein expression, this field is still far for the generation of a universal protocol, so many different parameters are necessary to be evaluated for each target. The development of robotic technologies has facilitated the evaluation of an important number of different conditions reducing cost and effort through miniaturization of experiments. At present there are novel technological approaches (strain engineering, fusion technologies, and protein purification, among others), which are key factors that should be used in the lab to increase the success for the production of a soluble and homogeneous target protein.

Acknowledgments

This work was financed by a research grant from FCE-7273 and FMV-7323, 2011 from Agencia Nacional de Investigación e Innovación (ANII), Montevideo, Uruguay to P. Oppezzo. A. Correa was financed by a doctoral fellowship from ANII, Uruguay.

References

1. Sorensen HP (2010) Towards universal systems for recombinant gene expression. Microb Cell Fact 9:27
2. Huang CJ, Lin H, Yang X (2012) Industrial production of recombinant therapeutics in Escherichia coli and its recent advancements. J Ind Microbiol Biotechnol 39:383–399
3. Yang Z, Zhang L, Zhang Y et al (2011) Highly efficient production of soluble proteins from insoluble inclusion bodies by a two-step-denaturing and refolding method. PLoS One 6:e22981
4. Correa A, Oppezzo P (2011) Tuning different expression parameters to achieve soluble recombinant proteins in E. coli: advantages of high-throughput screening. Biotechnol J 6:715–730
5. Samuelson JC (2011) Recent developments in difficult protein expression: a guide to E. coli strains, promoters, and relevant host mutations. Methods Mol Biol 705:195–209
6. Vincentelli R, Cimino A, Geerlof A et al (2011) High-throughput protein expression screening and purification in Escherichia coli. Methods 55:65–72
7. Vincentelli R, Canaan S, Campanacci V et al (2004) High-throughput automated refolding screening of inclusion bodies. Protein Sci 13:2782–2792
8. Foit L, Morgan GJ, Kern MJ et al (2009) Optimizing protein stability in vivo. Mol Cell 36:861–871
9. Hart DJ, Waldo GS (2013) Library methods for structural biology of challenging proteins and their complexes. Curr Opin Struct Biol 23:403–408
10. Artimo P, Jonnalagedda M, Arnold K et al (2012) ExPASy: SIB bioinformatics resource portal. Nucleic Acids Res 40:W597–W603
11. Gustafsson C, Govindarajan S, Minshull J (2004) Codon bias and heterologous protein expression. Trends Biotechnol 22: 346–353
12. Puigbo P, Guzman E, Romeu A et al (2007) OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. Nucleic Acids Res 35:W126–W131
13. Villalobos A, Ness JE, Gustafsson C et al (2006) Gene Designer: a synthetic biology tool for constructing artificial DNA segments. BMC Bioinformatics 7:285
14. Chung BK, Lee DY (2012) Computational codon optimization of synthetic gene for protein expression. BMC Syst Biol 6:134

15. Burgess-Brown NA, Sharma S, Sobott F et al (2008) Codon optimization can improve expression of human genes in Escherichia coli: a multi-gene study. Protein Expr Purif 59:94–102
16. Tegel H, Tourle S, Ottosson J et al (2010) Increased levels of recombinant human proteins with the Escherichia coli strain Rosetta(DE3). Protein Expr Purif 69: 159–167
17. Rosano GL, Ceccarelli EA (2009) Rare codon content affects the solubility of recombinant proteins in a codon bias-adjusted Escherichia coli strain. Microb Cell Fact 8:41
18. Marin M (2008) Folding at the rhythm of the rare codon beat. Biotechnol J 3:1047–1057
19. Voges D, Watzele M, Nemetz C et al (2004) Analyzing and enhancing mRNA translational efficiency in an Escherichia coli in vitro expression system. Biochem Biophys Res Commun 318:601–614
20. Salis HM, Mirsky EA, Voigt CA (2009) Automated design of synthetic ribosome binding sites to control protein expression. Nat Biotechnol 27:946–950
21. Salis HM (2011) The ribosome binding site calculator. Methods Enzymol 498:19–42
22. Makino T, Skretas G, Georgiou G (2011) Strain engineering for improved expression of recombinant proteins in bacteria. Microb Cell Fact 10:32
23. Salinas G, Pellizza L, Margenat M et al (2011) Tuned Escherichia coli as a host for the expression of disulfide-rich proteins. Biotechnol J 6:686–699
24. Ferre F, Clote P (2005) DiANNA: a web server for disulfide connectivity prediction. Nucleic Acids Res 33:W230–W232
25. Lin HH, Tseng LY (2010) DBCP: a web server for disulfide bonding connectivity pattern prediction without the prior knowledge of the bonding state of cysteines. Nucleic Acids Res 38:W503–W507
26. Berkmen M (2012) Production of disulfide-bonded proteins in Escherichia coli. Protein Expr Purif 82:240–251
27. Klint JK, Senff S, Saez NJ et al (2013) Production of recombinant disulfide-rich venom peptides for structural and functional analysis via expression in the periplasm of E. coli. PLoS One 8:e63865
28. Mergulhao FJ, Summers DK, Monteiro GA (2005) Recombinant protein secretion in Escherichia coli. Biotechnol Adv 23:177–202
29. den Blaauwen T, Driessen AJ (1996) Sec-dependent preprotein translocation in bacteria. Arch Microbiol 165:1–8
30. Luirink J, Sinning I (2004) SRP-mediated protein targeting: structure and function revisited. Biochim Biophys Acta 1694:17–35
31. Natale P, Bruser T, Driessen AJ (2008) Sec- and Tat-mediated protein secretion across the bacterial cytoplasmic membrane—distinct translocases and mechanisms. Biochim Biophys Acta 1778:1735–1756
32. Wagner S, Klepsch MM, Schlegel S et al (2008) Tuning Escherichia coli for membrane protein overexpression. Proc Natl Acad Sci U S A 105:14371–14376
33. Schlegel S, Rujas E, Ytterberg AJ et al (2013) Optimizing heterologous protein production in the periplasm of E coli by regulating gene expression levels. Microb Cell Fact 12:24
34. de Marco A (2009) Strategies for successful recombinant expression of disulfide bond-dependent proteins in Escherichia coli. Microb Cell Fact 8:26
35. Bessette PH, Aslund F, Beckwith J et al (1999) Efficient folding of proteins with multiple disulfide bonds in the Escherichia coli cytoplasm. Proc Natl Acad Sci U S A 96: 13703–13708
36. Lobstein J, Emrich CA, Jeans C et al (2012) SHuffle, a novel Escherichia coli protein expression strain capable of correctly folding disulfide bonded proteins in its cytoplasm. Microb Cell Fact 11:56
37. Hatahet F, Nguyen VD, Salo KE et al (2010) Disruption of reducing pathways is not essential for efficient disulfide bond formation in the cytoplasm of E. coli. Microb Cell Fact 9:67
38. Nguyen VD, Hatahet F, Salo KE et al (2010) Pre-expression of a sulfhydryl oxidase significantly increases the yields of eukaryotic disulfide bond containing proteins expressed in the cytoplasm of E. coli. Microb Cell Fact 10:1
39. Nozach H, Fruchart-Gaillard C, Fenaille F et al (2013) High throughput screening identifies disulfide isomerase DsbC as a very efficient partner for recombinant expression of small disulfide-rich proteins in E. coli. Microb Cell Fact 12:37
40. Walls D, Loughran ST (2011) Tagging recombinant proteins to enhance solubility and aid purification. Methods Mol Biol 681:151–175
41. Young CL, Britton ZT, Robinson AS (2012) Recombinant protein expression and purification: a comprehensive review of affinity tags and microbial applications. Biotechnol J 7:620–634
42. Murphy MB, Doyle SA (2005) High-throughput purification of hexahistidine-tagged proteins expressed in E. coli. Methods Mol Biol 310:123–130

43. Zhu XQ, Li SX, He HJ et al (2005) On-column refolding of an insoluble His6-tagged recombinant EC-SOD overexpressed in Escherichia coli. Acta Biochim Biophys Sin (Shanghai) 37:265–269
44. Li M, Su ZG, Janson JC (2004) In vitro protein refolding by chromatographic procedures. Protein Expr Purif 33:1–10
45. Schafer F, Romer U, Emmerlich M et al (2002) Automated high-throughput purification of 6xHis-tagged proteins. J Biomol Tech 13:131–142
46. Vincentelli R, Canaan S, Offant J et al (2005) Automated expression and solubility screening of His-tagged proteins in 96-well format. Anal Biochem 346:77–84
47. Steen J, Uhlen M, Hober S et al (2006) High-throughput protein purification using an automated set-up for high-yield affinity chromatography. Protein Expr Purif 46:173–178
48. Magnusdottir A, Johansson I, Dahlgren LG et al (2009) Enabling IMAC purification of low abundance recombinant proteins from E. coli lysates. Nat Methods 6:477–478
49. Bolanos-Garcia VM, Davies OR (2006) Structural analysis and classification of native proteins from E. coli commonly co-purified by immobilised metal affinity chromatography. Biochim Biophys Acta 1760:1304–1313
50. Robichon C, Luo J, Causey TB et al (2011) Engineering Escherichia coli BL21(DE3) derivative strains to minimize E. coli protein contamination after purification by immobilized metal affinity chromatography. Appl Environ Microbiol 77:4634–4646
51. Andersen KR, Leksa NC, Schwartz TU (2013) Optimized E. coli expression strain LOBSTR eliminates common contaminants from His-tag purification. Proteins 81:1857–1861
52. Schmidt TG, Skerra A (2007) The Strep-tag system for one-step purification and high-affinity detection or capturing of proteins. Nat Protoc 2:1528–1535
53. Lichty JJ, Malecki JL, Agnew HD et al (2005) Comparison of affinity tags for protein purification. Protein Expr Purif 41:98–105
54. Schmidt TG, Batz L, Bonet L et al (2013) Development of the Twin-Strep-tag(R) and its application for purification of recombinant proteins from cell culture supernatants. Protein Expr Purif 92:54–61
55. Hammarstrom M, Hellgren N, van Den Berg S et al (2002) Rapid screening for improved solubility of small human proteins produced as fusion proteins in Escherichia coli. Protein Sci 11:313–321
56. Esposito D, Chatterjee DK (2006) Enhancement of soluble protein expression through the use of fusion tags. Curr Opin Biotechnol 17:353–358
57. Pattenden LK, Thomas WG (2008) Amylose affinity chromatography of maltose-binding protein: purification by both native and novel matrix-assisted dialysis refolding methods. Methods Mol Biol 421:169–189
58. Smith DB, Johnson KS (1988) Single-step purification of polypeptides expressed in Escherichia coli as fusions with glutathione S-transferase. Gene 67:31–40
59. Dyson MR, Shadbolt SP, Vincent KJ et al (2004) Production of soluble mammalian proteins in Escherichia coli: identification of protein features that correlate with successful expression. BMC Biotechnol 4:32
60. Kapust RB, Waugh DS (1999) Escherichia coli maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. Protein Sci 8:1668–1674
61. Cho HJ, Lee Y, Chang RS et al (2008) Maltose binding protein facilitates high-level expression and functional purification of the chemokines RANTES and SDF-1alpha from Escherichia coli. Protein Expr Purif 60:37–45
62. LaVallie ER, Lu Z, Diblasio-Smith EA et al (2000) Thioredoxin as a fusion partner for production of soluble recombinant proteins in Escherichia coli. Methods Enzymol 326:322–340
63. Kim S, Lee SB (2008) Soluble expression of archaeal proteins in Escherichia coli by using fusion-partners. Protein Expr Purif 62: 116–119
64. LaVallie ER, DiBlasio EA, Kovacic S et al (1993) A thioredoxin gene fusion expression system that circumvents inclusion body formation in the E. coli cytoplasm. Biotechnology (N Y) 11:187–193
65. Marblestone JG, Edavettal SC, Lim Y et al (2006) Comparison of SUMO fusion technology with traditional gene fusion systems: enhanced expression and solubility with SUMO. Protein Sci 15:182–189
66. Malakhov MP, Mattern MR, Malakhova OA et al (2004) SUMO fusions and SUMO-specific protease for efficient expression and purification of proteins. J Struct Funct Genomics 5:75–86
67. Butt TR, Edavettal SC, Hall JP et al (2005) SUMO fusion technology for difficult-to-express proteins. Protein Expr Purif 43:1–9
68. Zhang Z, Li ZH, Wang F et al (2002) Overexpression of DsbC and DsbG markedly improves soluble and functional expression of single-chain Fv antibodies in Escherichia coli. Protein Expr Purif 26:218–228

69. De Marco V, Stier G, Blandin S et al (2004) The solubility and stability of recombinant proteins are increased by their fusion to NusA. Biochem Biophys Res Commun 322:766–771
70. Nallamsetty S, Waugh DS (2006) Solubility-enhancing proteins MBP and NusA play a passive role in the folding of their fusion partners. Protein Expr Purif 45:175–182
71. van den Berg S, Lofdahl PA, Hard T et al (2006) Improved solubility of TEV protease by directed evolution. J Biotechnol 121:291–298
72. Kapust RB, Tozser J, Copeland TD et al (2002) The P1′ specificity of tobacco etch virus protease. Biochem Biophys Res Commun 294:949–955
73. Moon AF, Mueller GA, Zhong X et al (2010) A synergistic approach to protein crystallization: combination of a fixed-arm carrier with surface entropy reduction. Protein Sci 19:901–913
74. Suzuki N, Hiraki M, Yamada Y et al (2010) Crystallization of small proteins assisted by green fluorescent protein. Acta Crystallogr D Biol Crystallogr 66:1059–1066
75. Smyth DR, Mrozkiewicz MK, McGrath WJ et al (2003) Crystal structures of fusion proteins with large-affinity tags. Protein Sci 12:1313–1322
76. Corsini L, Hothorn M, Scheffzek K et al (2008) Thioredoxin as a fusion tag for carrier-driven crystallization. Protein Sci 17: 2070–2079
77. Esposito D, Garvey LA, Chakiath CS (2009) Gateway cloning for protein expression. Methods Mol Biol 498:31–54
78. Berrow NS, Alderton D, Sainsbury S et al (2007) A versatile ligation-independent cloning method suitable for high-throughput expression screening applications. Nucleic Acids Res 35:e45
79. Unger T, Jacobovitch Y, Dantes A et al (2010) Applications of the Restriction Free (RF) cloning procedure for molecular manipulations and protein expression. J Struct Biol 172:34–44
80. Correa A, Ortega C, Obal G, Alzari P, Vincentelli R, Oppezzo P (2014) Generation of a vector suite for protein solubility screening. Front Microbiol. 5: 67
81. Erijman A, Dantes A, Bernheim R et al (2011) Transfer-PCR (TPCR): a highway for DNA cloning and protein engineering. J Struct Biol 175:171–177
82. Bond SR, Naus CC (2012) RF-Cloning.org: an online tool for the design of restriction-free cloning projects. Nucleic Acids Res 40:W209–W213
83. Vera A, Gonzalez-Montalban N, Aris A et al (2007) The conformational quality of insoluble recombinant proteins is enhanced at low growth temperatures. Biotechnol Bioeng 96:1101–1106
84. Studier FW (2005) Protein production by auto-induction in high density shaking cultures. Protein Expr Purif 41:207–234
85. Blommel PG, Becker KJ, Duvnjak P et al (2007) Enhanced bacterial protein expression during auto-induction obtained by alteration of lac repressor dosage and medium composition. Biotechnol Prog 23:585–598
86. Ukkonen K, Mayer S, Vasala A et al (2013) Use of slow glucose feeding as supporting carbon source in lactose autoinduction medium improves the robustness of protein expression at different aeration conditions. Protein Expr Purif 91:147–154
87. Krause M, Ukkonen K, Haataja T et al (2010) A novel fed-batch based cultivation method provides high cell-density and improves yield of soluble recombinant proteins in shaken cultures. Microb Cell Fact 9:11
88. Vincentelli R, Romier C (2013) Expression in Escherichia coli: becoming faster and more complex. Curr Opin Struct Biol 23:326–334
89. Koehn J, Hunt I (2009) High-throughput protein production (HTPP): a review of enabling technologies to expedite protein production. Methods Mol Biol 498:1–18
90. Ventura S, Villaverde A (2006) Protein quality in bacterial inclusion bodies. Trends Biotechnol 24:179–185
91. Dechavanne V, Barrillat N, Borlat F et al (2010) A high-throughput protein refolding screen in 96-well format combined with design of experiments to optimize the refolding conditions. Protein Expr Purif 75:192–203
92. Clark EDB (1998) Refolding of recombinant proteins. Curr Opin Biotechnol 9:157–163
93. Achmuller C, Kaar W, Ahrer K et al (2007) N(pro) fusion technology to produce proteins with authentic N termini in E. coli. Nat Methods 4:1037–1043
94. Ke T, Liang S, Huang J et al (2012) A novel PCR-based method for high throughput prokaryotic expression of antimicrobial peptide genes. BMC Biotechnol 12:10
95. Tokatlidis K, Dhurjati P, Millet J et al (1991) High activity of inclusion bodies formed in Escherichia coli overproducing Clostridium thermocellum endoglucanase D. FEBS Lett 282:205–208
96. Garcia-Fruitos E, Gonzalez-Montalban N, Morell M et al (2005) Aggregation as bacterial inclusion bodies does not imply inactiva-

tion of enzymes and fluorescent proteins. Microb Cell Fact 4:27

97. de Groot NS, Ventura S (2006) Protein activity in bacterial inclusion bodies correlates with predicted aggregation rates. J Biotechnol 125:110–113
98. Peternel S, Grdadolnik J, Gaberc-Porekar V et al (2008) Engineering inclusion bodies for non denaturing extraction of functional proteins. Microb Cell Fact 7:34
99. Garcia-Fruitos E (2010) Inclusion bodies: a new concept. Microb Cell Fact 9:80
100. Garcia-Fruitos E, Vazquez E, Diez-Gil C et al (2012) Bacterial inclusion bodies: making gold from waste. Trends Biotechnol 30:65–70
101. Villaverde A, Garcia-Fruitos E, Rinas U et al (2012) Packaging protein drugs as bacterial inclusion bodies for therapeutic applications. Microb Cell Fact 11:76
102. Low C, Moberg P, Quistgaard EM et al (2013) High-throughput analytical gel filtration screening of integral membrane proteins for structural studies. Biochim Biophys Acta 1830:3497–3508
103. Sala E, de Marco A (2010) Screening optimized protein purification protocols by coupling small-scale expression and mini-size exclusion chromatography. Protein Expr Purif 74:231–235
104. Hattori M, Hibbs RE, Gouaux E (2012) A fluorescence-detection size-exclusion chromatography-based thermostability assay for membrane protein precrystallization screening. Structure 20:1293–1299
105. Kawate T, Gouaux E (2006) Fluorescence-detection size-exclusion chromatography for precrystallization screening of integral membrane proteins. Structure 14:673–681
106. Backmark AE, Olivier N, Snijder A et al (2013) Fluorescent probe for high-throughput screening of membrane protein expression. Protein Sci 22:1124–1132
107. Dale GE, Broger C, Langen H et al (1994) Improving protein solubility through rationally designed amino acid replacements: solubilization of the trimethoprim-resistant type S1 dihydrofolate reductase. Protein Eng 7:933–939
108. Eijsink VG, Bjork A, Gaseidnes S et al (2004) Rational engineering of enzyme stability. J Biotechnol 113:105–120
109. Rasila TS, Pajunen MI, Savilahti H (2009) Critical evaluation of random mutagenesis by error-prone polymerase chain reaction protocols, Escherichia coli mutator strain, and hydroxylamine treatment. Anal Biochem 388:71–80
110. Stemmer WP (1994) Rapid evolution of a protein in vitro by DNA shuffling. Nature 370:389–391
111. Roodveldt C, Aharoni A, Tawfik DS (2005) Directed evolution of proteins for heterologous expression and stability. Curr Opin Struct Biol 15:50–56
112. Waldo GS, Standish BM, Berendzen J et al (1999) Rapid protein-folding assay using green fluorescent protein. Nat Biotechnol 17:691–695
113. Pedelacq JD, Piltch E, Liong EC et al (2002) Engineering soluble proteins for structural genomics. Nat Biotechnol 20:927–932
114. Cabantous S, Terwilliger TC, Waldo GS (2005) Protein tagging and detection with engineered self-assembling fragments of green fluorescent protein. Nat Biotechnol 23:102–107
115. Pedelacq JD, Cabantous S, Tran T et al (2006) Engineering and characterization of a superfolder green fluorescent protein. Nat Biotechnol 24:79–88
116. Maxwell KL, Mittermaier AK, Forman-Kay JD et al (1999) A simple in vivo assay for increased protein solubility. Protein Sci 8:1908–1911
117. Sieber V, Martinez CA, Arnold FH (2001) Libraries of hybrid proteins from distantly related sequences. Nat Biotechnol 19: 456–460
118. Dahlroth SL, Nordlund P, Cornvik T (2006) Colony filtration blotting for screening soluble expression in Escherichia coli. Nat Protoc 1:253–258
119. Cornvik T, Dahlroth SL, Magnusdottir A et al (2005) Colony filtration blot: a new screening method for soluble protein expression in Escherichia coli. Nat Methods 2: 507–509