UNIVERSIDAD DE GRANADA

FACULTAD DE CIENCIAS

DEPARTAMENTO DE QUÍMICA FÍSICA

BIBLIOTECAS COMBINATORIALES ENFOCADAS EN INGENIERÍA DE PROTEÍNAS: APLICACIONES EN TERMOESTABILIZACIÓN, PROMISCUIDAD Y DISEÑO DE SITIOS ACTIVOS

Héctor García Seisdedos

TESIS DOCTORAL
PROGRAMA DE DOCTORADO EN QUÍMICA
GRANADA
2012

UNIVERSIDAD DE GRANADA

FACULTAD DE CIENCIAS

DEPARTAMENTO DE QUÍMICA FÍSICA

BIBLIOTECAS COMBINATORIALES ENFOCADAS EN INGENIERÍA DE PROTEÍNAS: APLICACIONES EN TERMOESTABILIZACIÓN, PROMISCUIDAD Y DISEÑOS DE SITIOS ACTIVOS

Héctor García Seisdedos

TESIS DOCTORAL
PROGRAMA DE DOCTORADO EN QUÍMICA
GRANADA
2012

Editor: Editorial de la Universidad de Granada
Autor: Héctor García Seisdedos
D.L.: GR 508-2013
ISBN: 978-84-9028-389-9

UNIVERSIDAD DE GRANADA
FACULTAD DE CIENCIAS
DEPARTAMENTO DE QUÍMICA FÍSICA

Memoria presentada para optar al Grado de Doctor por la Universidad de Granada

Fdo.: Héctor García Seisdedos
Licenciado en Ciencias Biológicas y Licenciado en Bioquímica por la Universidad de Salamanca

Granada Julio 2012

Directores de la Tesis:

Fdo.: José Manuel Sánchez Ruiz
Catedrático de Química Física
Departamento de Química Física
Facultad de Ciencias
Universidad de Granada

Fdo.: Beatriz Ibarra Molero
Profesora Titular
Departamento de Química Física
Facultad de Ciencias
Universidad de Granada

Los directores de esta Tesis Doctoral, José Manuel Sánchez Ruiz y Beatriz Ibarra Molero y el doctorando, Héctor García Seisdedos,
garantizamos al firmar la Tesis Doctoral, que el trabajo ha sido realizado por el doctorando bajo nuestra dirección, y hasta donde nuestro conocimiento alcanza, en la realización del trabajo se han respetado los derechos de otros autores, al ser citados cuando se han utilizado sus resultados o publicaciones.

Julio 2012 Granada

Directores de la Tesis:

Fdo.: José Manuel Sánchez Ruiz Fdo.: Beatriz Ibarra Molero

Doctorando:

Fdo.: Héctor García Seisdedos

Yo, José Manuel Sánchez Ruiz, con DNI nº 24141354-W. Siendo coautor de la publicación *"Probing the Mutational Interplay between Primary and Promiscuous Protein Functions: A Computational-Experimental Approach"* publicada en la revista PLoS Computational Biology con fecha 14 de junio de 2012.

Declaro que:

- Héctor García Seisdedos, el también coautor de esta publicación, realizó los experimentos y analizó los datos que se presentan en esta publicación.
- No he presentado y renuncio a presentar dicha publicación en otra Tesis Doctoral.

En Granada, con fecha 12 de julio.

Fdo: José Manuel Sánchez Ruiz
Catedrático de Química-Física
Departamento de Química-Física
Facultad de Ciencias
Universidad de Granada

Yo, Beatriz Ibarra Molero, con DNI nº 24256088-N. Siendo coautora de la publicación *"How many ionizable groups can sit on a protein hydrophobic core?"* publicada en la revista Proteins Structure, Function and Bioinformatics, con fecha 30 de septiembre de 2011.

Declaro que:

- Héctor García Seisdedos, el también coautor de esta publicación, realizó los experimentos y analizó los datos que se presentan en esta publicación.
- No he presentado y renuncio a presentar dicha publicación en otra Tesis Doctoral.

En Granada, con fecha 12 de julio.

Fdo: Beatriz Ibarra Molero

Profesora Titular

Departamento de Química-Física

Facultad de Ciencias

Universidad de Granada

Yo, José Manuel Sánchez Ruiz, con DNI nº 24141354-W. Siendo coautor de la publicación *"How many ionizable groups can sit on a protein hydrophobic core?"* publicada en la revista Proteins Structure, Function and Bioinformatics, con fecha 30 de septiembre de 2011.

Declaro que:

- Héctor García Seisdedos, el también coautor de esta publicación, realizó los experimentos y analizó los datos que se presentan en esta publicación.
- No he presentado y renuncio a presentar dicha publicación en otra Tesis Doctoral.

En Granada, con fecha 12 de julio.

Fdo: José Manuel Sánchez Ruiz
Catedrático de Química-Física
Departamento de Química-Física
Facultad de Ciencias
Universidad de Granada

A mis padres

A mis hermanos

A Sisi

Agradecimientos

Agradezco, sinceramente, a Beatriz y a José Manuel, en primer lugar haberme dado la oportunidad de realizar esta tesis doctoral, ellos me han enseñado prácticamente todo lo que he aprendido en estos años. Gracias por vuestro entusiasmo y dedicación. Soy consciente de que he tenido mucha suerte.

En segundo lugar agradecerle de manera muy especial a Sisi por muchas razones. Sisi, ha sido muy bonito vernos madurar durante estos años, tanto profesional como personalmente, y me hace muy feliz tenerte a mi lado.

A la gente del departamento, en especial a María del Mar, a Mari Carmen, a Asun y a Israel por su amistad y los buenos momentos vividos también fuera del trabajo. Y A mis compañeros de laboratorio: Inma, Álvaro, Noel, Ángel, Valeria, Diego y Rocío.

A mis compañeros del laboratorio de Paris.

A todos mis amigos de Ávila, Salamanca y Granada porque siempre que lo he necesitado habéis estado ahí. ¡Gracias!

Por último agradecer a mi familia, que me lo ha dado todo, el apoyo y comprensión que muestran siempre. Estoy muy contento de poder agradecérselo aquí.

Índice

1 Introducción y objetivos

Las proteínas son las macromoléculas más versátiles de entre las que forman los organismos vivos, no en vano constituyen alrededor del 50% del peso seco de las células y tejidos, y es que forman parte fundamental de su arquitectura e intervienen de manera crucial en casi cualquier proceso biológico. Pueden ser consideradas como las "moléculas trabajadoras" de las células, ya que catalizan un extraordinario rango de reacciones químicas, son responsables de la generación de energía en las células (fotosíntesis y respiración celular), transportan y almacenan otras moléculas como el oxigeno, generan movimiento, actúan de mensajeras transmitiendo señales que coordinan procesos en y entre células, tejidos y órganos, controlan el crecimiento y la diferenciación de los organismos, proporcionan protección inmunológica (anticuerpos), y un largo etcétera de funciones, incluso la de bioluminiscencia (ver Figura 1). Se podría decir que son el brazo efector de la información genética, por tanto se hace imprescindible su estudio en cualquier intento de comprensión de un determinado mecanismo biológico.

En los últimos años con el desarrollo de los campos de la biotecnología y la biomedicina ha ido generalizándose el uso de las proteínas a nivel tecnológico, usando su eficiente poder catalizador en multitud de procesos industriales que van desde la industria textil y alimenticia hasta la cosmética y farmacéutica, y es que las enzimas son los más eficientes catalizadores de reacciones químicas conocidos, llegando a aumentar la velocidad de una reacción en hasta 23 órdenes de magnitud [1, 2]. También son utilizadas como biosensores, en biorremediación (enzimas que degradan productos contaminantes), marcadores de procesos celulares (ver Figura 1), así como en multitud de aplicaciones terapéuticas, por ejemplo anticuerpos monoclonales, insulina, hormona de crecimiento, interferón…

Para el uso aplicado de las proteínas, uso que normalmente supone sacarlas del contexto natural en el que se hayan ubicadas y para el que han evolucionado, casi siempre se hacen necesarias ciertas modificaciones con

objeto de optimizarlas respecto a alguna propiedad importante para permitir un uso eficiente de las mismas. Un ejemplo claro lo representa la estabilidad de la que se sabe que suele ser marginal, lo cual tiene su razón de ser *in vivo,* pero que para su uso en laboratorio o en la industria puede ser un problema [3]; otras propiedades importantes a optimizar podrían ser la solubilidad, la actividad enzimática en determinadas condiciones de trabajo, la eliminación de potenciales problemas de inmunogenicidad o toxicidad en el caso de proteínas terapéuticas, etcétera.

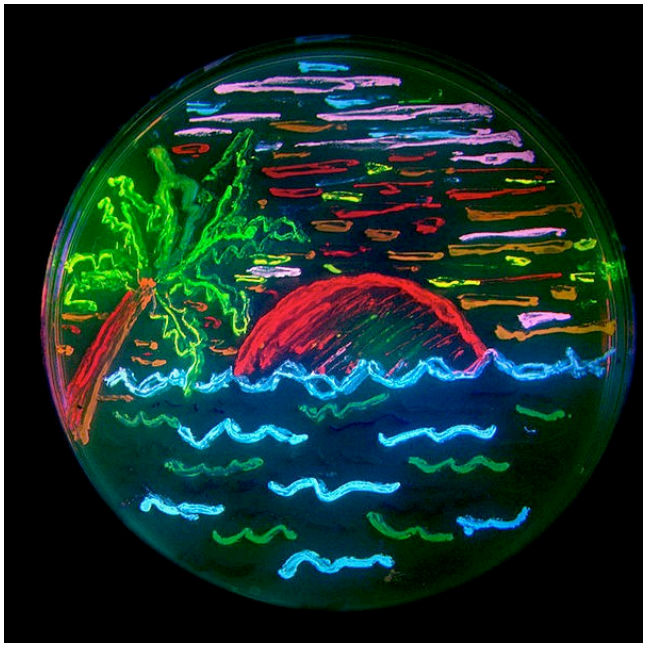

Figura 1. Placa Petri sembrada con bacterias que expresan diferentes versiones de las proteínas fluorescentes GFP y RFP. La GFP (proteína verde fluorescente) es producida por la medusa *Aequorea victoria* que emite bioluminiscencia en la zona verde del espectro visible, esta proteína ha sido ampliamente utilizada desde su descubrimiento en 1962 como marcador de procesos celulares. Los Doctores Shimomura, Chalfie y Tsien recibieron el Premio Nobel de Química en el 2008 por los descubrimientos relacionados con la GFP. Posteriormente fueron descubiertas más proteínas fluorescentes como la RFP (proteína roja fluorescente) de *Discosoma coral*. El grupo de investigación de Royer Y. Tsien de la UCSD entre otros ha modificado ambas proteínas consiguiendo variantes que emiten luz en varias longitudes de onda y así extendiendo la gama de marcadores de manera que se puedan usar para estudiar varios procesos de forma simultánea. Imagen tomada de **http://www.tsienlab.ucsd.edu/Images.htm**

Sin embargo, un objetivo mucho más ambicioso viene dado por el desarrollo de estrategias que permitan diseñar proteínas con actividades catalíticas nuevas, lo que implica el diseño de nuevos centros activos [4-11], que adquieran nuevas estructuras no presentes en la naturaleza [12], o bien aumentar actividades residuales ya existentes en las proteínas, llamadas actividades promiscuas, para utilizarlas con fines biotecnológicos [11, 13-21]. De todo esto se encarga el campo de la ingeniería de proteínas.

Desde un punto de vista metodológico, en los últimos años se han venido empleando tres enfoques en la ingeniería de proteínas:

El **diseño racional** o computacional, que valiéndose del conocimiento acumulados sobre la relación estructural-energética que determina las propiedades de las proteínas, trata de diseñarlas para mejorar dichas propiedades o para construir nuevas. Entre otras, algunas de las aportaciones del diseño racional a la ingeniería de proteínas han sido:

- Diseño de proteínas más estables, se ha demostrado que es posible aumentar la estabilidad de las proteínas rediseñando el núcleo hidrófobo [3, 22] o la distribución de cargas de la superficie [23-25], en esto último tiene una amplia experiencia nuestro grupo de investigación, incluyendo mutaciones del tipo (Xxx→Pro o Gly→Xxx) o introduciendo puentes disulfuro que actúen sobre la entropía configuracional, introduciendo puentes salinos, introduciendo residuos que interaccionan favorablemente con el dipolo de las α-hélices, etcétera [26].
- Diseño *de novo* de actividades catalíticas, por este procedimiento se han diseñado con éxito actividades catalíticas como la esterasa [5, 9], e incluso algunas actividades que no están presentes en ninguna de las proteínas conocidas y que catalizan célebres reacciones de la química orgánica como la eliminación de Kemp, retro aldólica y Diels Alder [6-8, 11]. Suponen el diseño de nuevos centros activos, modelando el

estado de transición de la unión con el sustrato en partes de proteínas que sean geométricamente compatibles con las del nuevo centro activo. El diseño de actividades catalíticas suelen venir de la mano de un fuerte efecto desestabilizante, por tanto han de ser diseñados en proteínas con suficiente estabilidad para sobrellevarlo, lo cual limita la búsqueda de proteínas con partes que tengan una geometría compatible a la del diseño.

- Diseño de receptores proteicos con el sitio activo modificado para unir moléculas diferentes a los ligandos naturales por ejemplo al TNT [27].
- Rediseño de interacciones proteína-ARN y proteína-ADN, que permitan el diseño de nuevas endonucleasas específicas [4, 10].

El grupo de David Baker de la Universidad de Washington, al que pertenecen gran parte de los logros citados, incluso ha conseguido diseñar una proteína, conocida como Top7, con un plegamiento *de novo*, no existente en la naturaleza [12].

Los trabajos publicados bajo este enfoque, aunque realmente espectaculares en algunos casos, siguen siendo muy escasos en número, y aún cuando se tiene éxito rara vez se consigue la misma eficacia que presentan las enzimas cuando catalizan su sustrato natural. El problema es que en la práctica, nuestros conocimientos acerca de los principios básicos que determinan la relación estructura/energética de las proteínas son aún bastante incompletos, por no hablar de que no siempre se dispone de la estructura de la proteína que se desea optimizar, lo cual es condición *sine qua non* para usar este enfoque. Por estas razones el enfoque racional-computacional aun presenta profundas limitaciones.

La alternativa es la llamada **evolución dirigida in vitro.** En esta se crea aleatoriamente una biblioteca de variantes de la proteína a optimizar y se analiza en términos de la característica deseada, con lo cual no se requiere apenas más información que la estructura primaria de la proteína. Se utiliza

un algoritmo basado en la evolución darwiniana (ver Figura 2); en un primer paso se genera variabilidad sobre el ADN de la proteína mediante mutagénesis aleatoria aplicando posteriormente una presión evolutiva, una selección, por la propiedad que se desea mejorar; las proteínas mejoradas serán las nuevas variantes de partida sobre las que realizar un nuevo ciclo, el algoritmo pararía cuando se hayan obtenido variantes con el grado de optimización deseado, o cuando, tras un número dado de ciclos, se observe que la propiedad no se puede optimizar más. Esta estrategia ha sido muy usada en la última década para aumentar la actividad catalítica, afinidad de unión a ligandos o la estabilidad de las proteínas [28-30], incluso ha llegado a combinarse con el diseño racional para el aumento de actividades *de novo* [8, 31, 32].

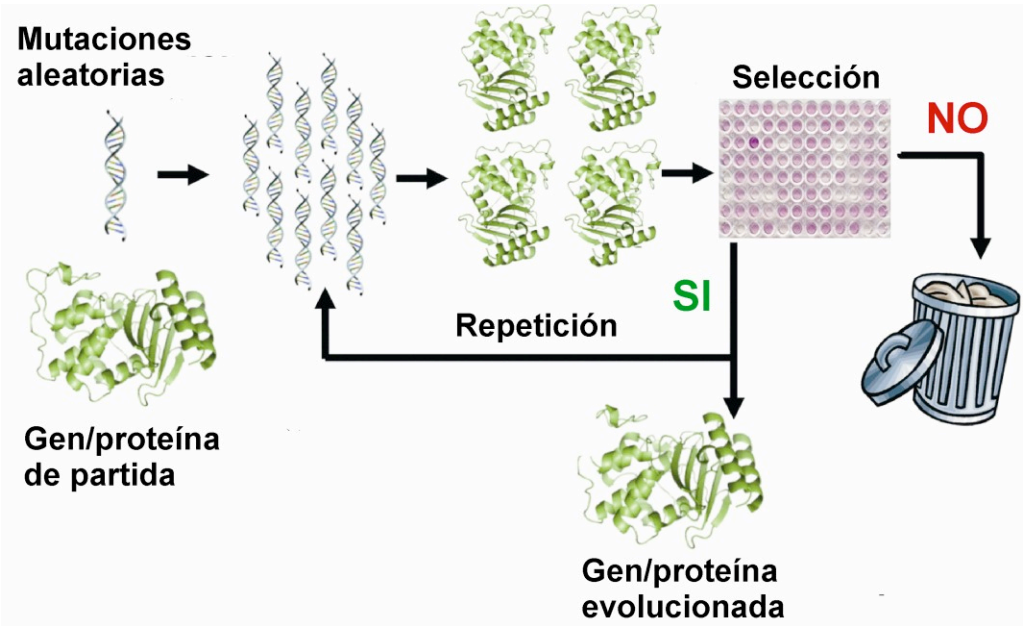

Figura 2. Esquema de un procedimiento típico de evolución dirigida *in vitro.* El Algoritmo parte de la proteína que se quiere optimizar por una determinada propiedad, en un primer paso se generan variantes de esta proteína por mutagénesis aleatoria, parte de las variantes (mayor o menor dependiendo de la eficacia del método de estudio) serán estudiadas por la propiedad a optimizar, seleccionando las mejores variantes para que actúen como "padres" en el siguiente ciclo de mutagénesis-selección, las variantes no seleccionadas son

descartadas. El algoritmo para cuando no sea posible aumentar más la propiedad a optimizar. Figura modificada de [13].

En principio esta estrategia funciona siempre y cuando se disponga de un método eficaz para estudiar un muy elevado número de variantes, ya que las bibliotecas de proteínas suelen ser muy grandes, incluso a veces del orden de millones de variantes [33], por tanto si no se dispone de un método eficaz de muestreo pueden no llegar a detectarse variantes interesantes [28].

Para dar solución a los problemas inherentes a la evolución dirigida, en los últimos años se ha venido desarrollando una tercera estrategia conocida ya como **diseño semi-racional**, esta comprendería parte de las dos estrategias anteriores, abordando el diseño de pequeñas bibliotecas de mutantes, en las cuales las mutaciones no serían aleatorias, si no que estarían propuestas bien por el diseño racional, bien por la información evolutiva (alineamientos de secuencias) o bien por cualquier otro criterio, por ejemplo estructural, que limite las dimensiones de la biblioteca. Podemos encontrar algunos ejemplos de ingeniería de proteínas usando este enfoque [33, 34], en concreto uno de ellos ha sido publicado recientemente por nuestro grupo de investigación, en el se diseñó una pequeña biblioteca combinatorial de variantes de tiorredoxina de *E. coli* con mutaciones propuestas por un análisis de consenso. Del estudio de un pequeño número, se encontraron algunas variantes, que resultaron ser optimizadas simultáneamente por varias propiedades (estabilidad térmica, cinética y actividad catalítica).

En la presente tesis doctoral, con título "Bibliotecas combinatoriales enfocadas en ingeniería de proteínas: Aplicaciones en termoestabilización, promiscuidad y diseño de sitios activos.", se presentan tres trabajos que se encuadran en esta tercera estrategia, dos de ellos han sido publicados en revistas internacionales de prestigio reconocido:

- Garcia-Seisdedos, H., B. Ibarra-Molero, and J.M. Sanchez-Ruiz, *How many ionizable groups can sit on a protein hydrophobic core?* Proteins, 2011. **80**(1): p. 1-7.
- Garcia-Seisdedos, H., B. Ibarra-Molero, and J.M. Sanchez-Ruiz, *Probing the Mutational Interplay between Primary and Promiscuous Protein Functions: A Computational-Experimental Approach.* PLoS Comput Biol, 2012. **8**(6): p. e1002558.

Un tercer trabajo se presenta en forma de borrador por encontrarse en vías de publicación (apartado 3.3). En estos tres trabajos tomando como proteína modelo la tiorredoxina de *E. coli*, se abordan diferentes objetivos claves en la ingeniería de proteínas, como el diseño de sitios activos, la estabilización térmica y el aumento de actividades promiscuas. Todos ellos supusieron el diseño de pequeñas bibliotecas combinatoriales de variantes cuyas mutaciones fueron propuestas:

- Por el diseño racional en el caso de la biblioteca dirigida a la termoestabilización (apartado 3.3).
- Por la información evolutiva de los alineamientos de secuencias en el caso de la biblioteca dirigida al aumento actividades promiscuas (apartado 3.2)
- Por información estructural en el caso de la biblioteca enfocada al diseño de centros activos. (apartado 3.1)

Para el estudio de estas bibliotecas combinatoriales se puso a punto un método en el cual se compagina el estudio experimental con una herramienta de regresión multivariante, el PLS-R (regresión por mínimos cuadrados parciales). Así del estudio experimental de unas pocas variantes se hace posible modelar la relación entre la secuencia y las propiedades estudiadas, y de esta forma poder llegar a predecir el comportamiento de todas las variantes que conforman la biblioteca. Por último, la verificación experimental de las mejores variantes predichas permitirá tener más datos

para volver a realizar otra predicción presumiblemente mejor que la primera. El proceso se puede ciclar de manera que con cada iteración se vaya explorando la biblioteca, en búsqueda de las mejores variantes, deteniendo el algoritmo cuando no se encuentre mejora significativa entre las predicciones de dos iteraciones consecutivas. En la Figura 3 se ilustra un esquema del proceso.

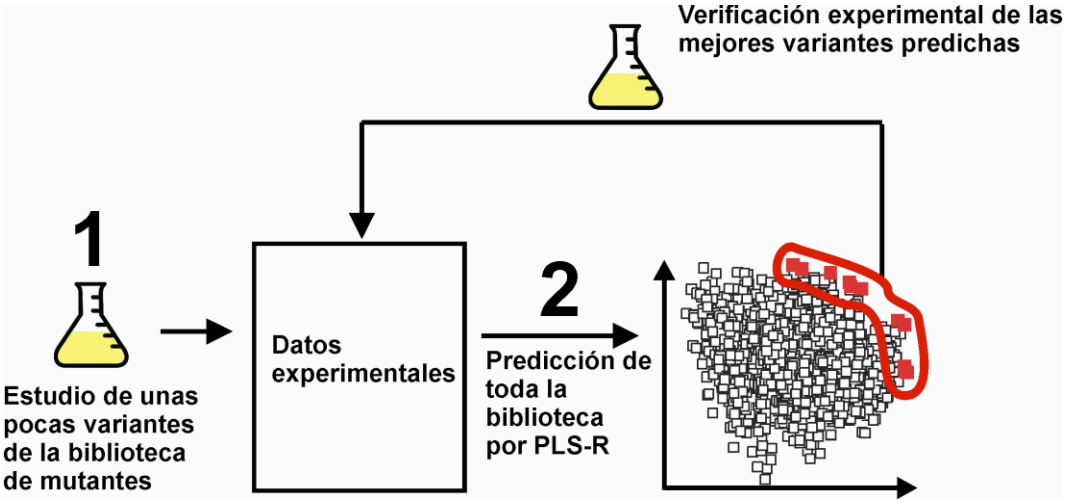

Figura 3. Esquema del algoritmo experimental-computacional desarrollado por nosotros para el estudio de bibliotecas de mutantes.

Esta metodología, descrita con más detalle en los apartados 2.5 y 3.2, fue aplicada al estudio de las tres bibliotecas combinatoriales diseñadas en este trabajo. En el caso del la biblioteca dirigida al aumento de actividades promiscuas, en la que se buscó el aumento simultaneo de varias propiedades, se utilizó el concepto de la frontera de Pareto para encontrar las mejores variantes. El concepto de la frontera de Pareto, explicado con detalle en los apartados 3.2 y 2.5, es familiar en el campo de las ciencias sociales y últimamente está empezando a extenderse su uso en campos como la biología evolutiva [35] y la ingeniería de proteínas [9, 36, 37]. Dicho concepto es aplicado a problemas de optimización multi-objetivo, donde no es posible encontrar una sola solución óptima, si no un conjunto de ellas, que conformarán la frontera de Pareto. Una solución A pertenecerá a dicho

conjunto, si y solo si, no existiera ninguna otra solución B que sea mejor que A en todos los objetivos tratados de forma simultánea.

El uso de métodos de regresión en el estudio de bibliotecas combinatoriales está comenzando a ser explorado [38], y a nuestro entender resulta una herramienta muy útil en el estudio de bibliotecas de mutantes permitiendo aumentar de forma considerable el tamaño de estas y por tanto su potencial.

2 Metodología

2.1 Obtención del ADN de proteínas mutantes

El ADN de los mutantes estudiados en este trabajo se ha obtenido por dos métodos. Unas variantes se generaron mediante mutaciones puntuales siguiendo el método que describiremos a continuación en el apartado 2.1.1, mientras que otras eran parte de bibliotecas combinatoriales de mutaciones, cuya construcción describiremos en el apartado 2.1.2. Todas las variantes estudiadas se construyeron sobre el vector pET30a (+), Figura 4. Este vector es un plásmido bacteriano caracterizado por tener un alto nivel de expresión. Dicha expresión está controlada por el promotor T7, lo que le hace dependiente de la T7 ARN polimerasa. El gen de tiorredoxina de *E. coli* se clonó entre las secuencias de corte de XhoI y NdeI. El vector también contiene un gen de resistencia a Kanamicina.

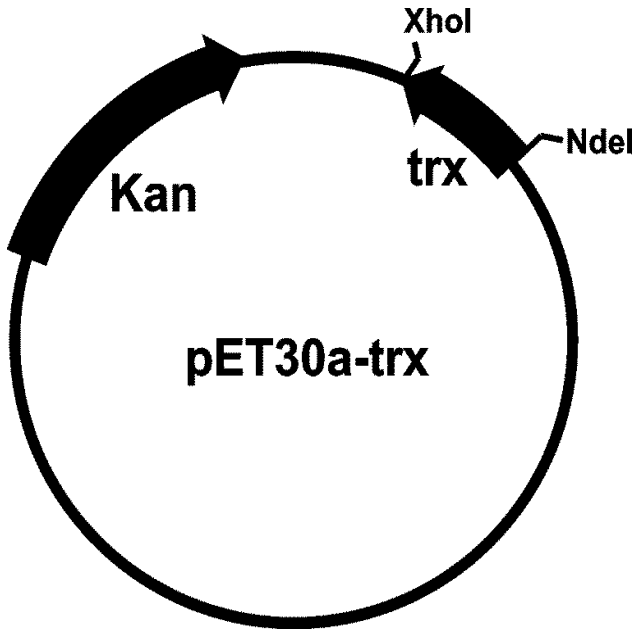

Figura 4. Mapa del plásmido pET30-trx. El gen que codifica para tiorredoxina (trx) ha sido clonado entre las secuencias de corte para Xhol y Ndel. La longitud total del plásmido es de aproximadamente 5600 pares de bases (pb).

El gen de la tiorredoxina de *E. coli* utilizado en este trabajo en realidad codifica una variante de la misma que contiene una cola de histidina, que

consta de 6 histidinas, en su extremo N terminal, con objeto de facilitar la purificación de la proteína.

2.1.1 Mutagénesis dirigida:

Para la realización de la mutagénesis dirigida utilizamos el QuickChange® II Site-Directed Mutagenesis Kit de Stratagene, este kit permite realizar mutaciones puntuales en un plásmido de doble cadena (ADN-dc). Este método es capaz de generar mutantes con una eficiencia mayor del 80% en un procedimiento que requerirá del (ADN-dc) molde que queramos mutar, que ha de estar metilado (provenir de una cepa de E coli dam^+), y de dos oligonucleótidos sintéticos, uno para cada cadena de ADN; estos oligonucleótidos contendrán la mutación deseada.

El procedimiento de obtención y purificación del ADN mutante constará de cinco pasos fundamentales, que vienen descritos en el kit comercial:

a. Diseño de oligonucleótidos, deben cumplir una serie de requisitos:

- Ambos deben contener la mutación preferiblemente en posición central.
- Deben tener una longitud de entre 25 y 45 pb, y una temperatura de fusión (T_m) ≥ 78 ^oC.
- Deben tener un contenido mínimo de bases GC del 40%.

Para el diseño de los oligonucleótidos nos ayudamos de la siguiente herramienta web: http://bioinformatics.org/primerx.

b. Síntesis del ADN mutante mediante PCR: el ADN mutante será sintetizado y amplificado por PCR a partir del ADN molde y de los dos cebadores (oligonucleótidos) sintetizados.

c. Digestión con Dpn I del ADN molde. El ADN del que hemos partido se encuentra metilado por proceder de una cepa de *E. coli* dam^+, sin embargo el ADN mutante generado en la PCR no lo estará. La DpnI es una endonucleasa específica para ADN metilado, de tal forma que esta

digerirá el ADN molde del producto de PCR, siendo el ADN mutante el único ADN que quedará en el producto de PCR.

d. Transformación del producto tratado con DpnI en células supercompetentes XL1-Blue, que tienen una alta capacidad de ser transformadas y que amplificaran el ADN mutado. Las células transformantes se sembrarán en placas petri con LB-agar suplementado con kanamicina a una concentración final de de 30µg/ml.

e. Extracción del ADN plasmídico mutante mediante el kit QIAprep®

Los protocolos no han sido descritos por haber seguido exactamente los especificados por los kits comerciales. Una vez obtenido parte del ADN mutante se secuencia y el resto se conserva a -20ºC.

2.1.2 Construcción de bibliotecas combinatoriales:

Llamaremos biblioteca o genoteca combinatorial al conjunto de variantes de un gen, en nuestro caso el de tiorredoxina, que contenga todas las combinaciones posibles de una serie de mutaciones elegidas *a priori*. Las bibliotecas combinatoriales usadas en este trabajo fueron construidas siguiendo el método descrito por Bessette, P.H., *et al*. [39], que se basa en el hecho de que se puedan construir genes enteros a partir de pequeños oligonucleótidos solapantes que pueden ser ensamblados por la ADN polimerasa [40]. Se sintetizan los oligos necesarios para construir todas las variantes del gen, y al irse éstos ensamblando darán lugar a todas las variantes de la biblioteca combinatorial.

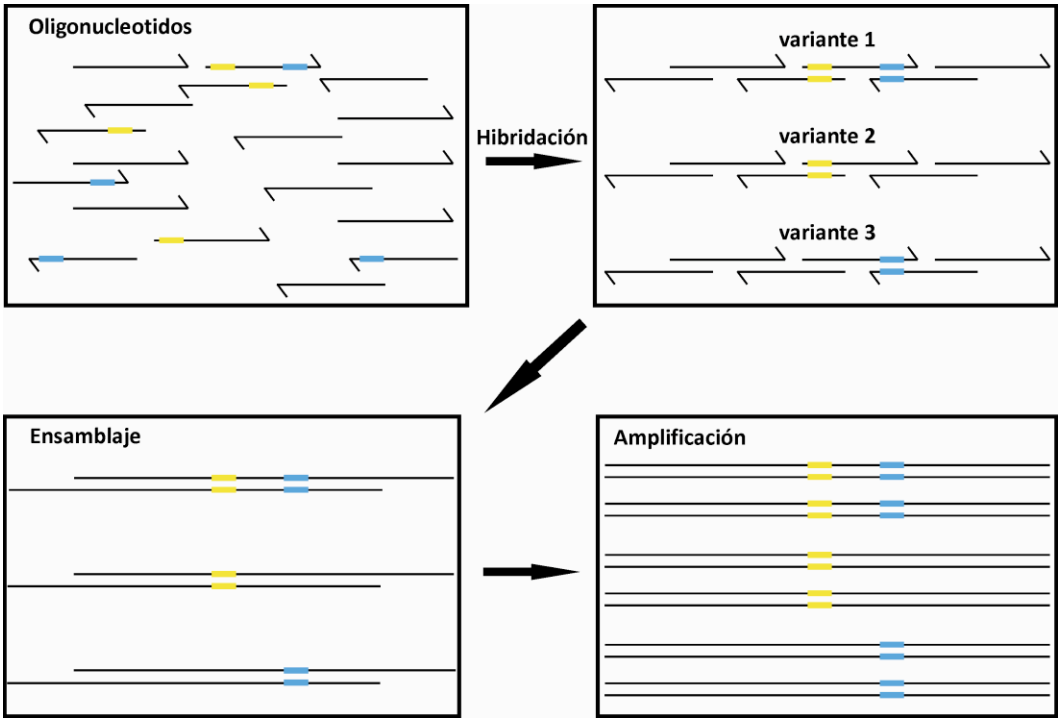

Figura 5. Ejemplo de la generación de una pequeña biblioteca combinatorial, los rectángulos amarillos y azules representan dos mutaciones puntuales. En primer lugar, hay que sintetizar los oligonucleótidos necesarios que, posteriormente, al hibridar y ensamblarse darán lugar a las distintas variantes que conforman la biblioteca. Posteriormente estas variantes serán amplificadas.

Este método consta de los siguientes pasos:

a. Diseño de los oligonucleótidos, estos deben tener una longitud de unos 40 pares de bases y han de codificar las dos hebras del ADN.

b. Ensamblaje por PCR.

c. Amplificación del producto por PCR, se añadirá al producto ensamblado un exceso de los oligonucleótidos de los extremos 5' de cada hebra, con objeto de que se amplifiquen las variantes del gen enteras.

d. Clonación del producto amplificado en nuestro vector (pET30a (+)).

Las principales ventajas de este método frente a otros de generación de bibliotecas combinatoriales es que, con este método nos aseguramos de que todas las mutaciones estén representadas, de que no haya limitación de

proximidad entre mutaciones (ya que varias mutaciones pueden estar dentro del mismo oligo) y de la ausencia de contaminación de ADN del gen silvestre en el producto final tan frecuente en otros métodos.

Cabe resaltar que frecuentemente encontramos variantes con mutaciones erróneas, es decir diferentes a las elegidas por nosotros, que simplemente desechamos.

También se ha observado experimentalmente que aproximadamente la mitad de las colonias transformantes con el ADN de la biblioteca no poseen el gen de la tiorredoxina, luego no pueden producir proteína.

2.1.3 Eliminación de la cola de histidina

La gran mayoría de las proteínas de este trabajo fueron estudiadas con la cola de histidina, pero en algún caso puntual se realizaron estudios de alguna variante sin cola de histidina. Para su eliminación se procedió como indicamos a continuación.

Lo que hicimos fue cortar el fragmento del gen de tiorredoxina sin la cola de histidina y clonarlo en otro vector pET30a (+). El procedimiento seguido fue el siguiente:

a. El primer paso fue amplificar por PCR el gen de tiorredoxina excluyendo la cola de histidina utilizando dos cebadores que además dotan a ambos extremos de sitios de corte para las enzimas NdeI y XhoI.

b. Se comprobó el fragmento amplificado mediante un gel de agarosa al 0.8%, y se purificó el producto de PCR valiéndonos de un MinElute® PCR Purification Kit de Qiagen.

c. Se procedió a la digestión, por separado, tanto del producto de PCR purificado (inserto) como del vector pET30a (+) mediante las enzimas NdeI y XhoI de New England Biolabs.

d. Purificamos ambos productos con el kit de purificación utilizado anteriormente y corrimos un gel de agarosa al 0.8% para cuantificar la cantidad de inserto y de vector que tenemos, en ng.

e. Se procedió a su ligación utilizando la ligasa de ADN T4, de New England Biolabs. Se añade 3 veces más inserto que vector. El volumen final fueron 10 µL y se incubó toda la noche a 16 ºC.

f. El producto de ligación es transformado en células supercompetentes XL-1 blue para amplificarlo.

g. Se extrae el ADN plasmídico de las células siguiendo el protocolo del kit QIAprep® (de Qiagen)

h. Por último, con objeto de comprobar que el proceso se ha completado correctamente, cortamos una pequeña parte del ADN purificado en el último paso y lo digerimos con las enzimas de restricción NdeI y XhoI. Tras la digestión corremos un gel de agarosa con el resultado de esta. Si el inserto, que codifica para la tiorredoxina de *E. coli* sin cola de histidina, se ha clonado correctamente en el vector, deberemos ver dos bandas en el gel, una de unos 5500 pb que corresponderá al vector y otra de unos 350 pb que corresponderá a nuestro inserto.

2.1.4 Secuenciación de ADN y síntesis de oligonucleótidos

La secuenciación de ADN y la síntesis de oligonucleótidos fueron llevadas a cabo por la empresa Eurofins MWG Operon (http://www.eurofinsdna.com). Las muestras enviadas para secuenciación contenían aproximadamente 1.5µg de ADN plasmídico liofilizado.

2.2 Obtención de proteínas

El procedimiento de obtención de la proteína para su estudio *in vitro* consta de los siguientes pasos:

2.2.1 Transformación del ADN plasmídico

El ADN que codifica para nuestra proteína se encuentra formando parte de un plásmido como describimos en el apartado anterior. Para obtener la proteína, el primer paso fue hacer la transformación de células competentes de la cepa BL21 (DE3) de Stratagene con dicho plásmido. Las células BL21 (DE3) se caracterizan por sus altos niveles de expresión y su fácil inducción, son usadas para la expresión de vectores que están bajo el control del promotor T7, es decir que necesitan de la T7 ARN polimerasa para su expresión; estas células tienen inserto en su cromosoma el ADN del fago λDE3 que codifica la T7 ARN polimerasa (bajo el control del promotor lacUV5) cuya expresión puede ser inducida por Isopropil-ß-D-tiogalactósido (IPTG).

El protocolo seguido para realizar la transformación fue el comercial.

2.2.2 Pruebas de expresión y criopreservación de células.

Tras la transformación de las células se procedió a realizar unas pruebas de expresión previas a la purificación de las proteínas, este paso es muy recomendable sobre todo en la purificación de proteínas de bibliotecas combinatoriales, ya que como mencionamos anteriormente aproximadamente la mitad de las colonias transformantes no poseen el inserto, es decir el gen la proteína a purificar. Por tanto mediante unas simples pruebas expresión podremos escoger las colonias transformantes que sobre-expresen la proteína de interés para su posterior purificación. Los pasos seguidos fueron:

a. Cada colonia transformante se inocula en 10 mL de medio de cultivo LB (Luria-Bertani) suplementado con kanamicina a una concentración final de 30µg/ml y se incuban en agitación a 37ºC.

b. Cuando el cultivo alcance una densidad óptica, absorbancia, de 0.6 a 600nm, es decir cuando se encuentre en fase exponencial de crecimiento:

I. Se recogen 0.5 mL de cada cultivo con objeto de preservar las células a -80ºC, se mezclarán bien con otros 0.5 mL de glicerol al 60% (esterilizado), que actuará de crioprotector, y se guardarán en un criotubo para almacenarlas a -80ºC para posteriores usos de las mismas.

II. Posteriormente inducimos los cultivos añadiendo IPTG hasta una concentración final de 0.4 mM y los dejamos en agitación a 37ºC de 6 a 18 horas.

c. Corremos una electroforesis SDS en un gel al 15% de poliacrilamida las muestras se prepararon mezclando 15µL del cultivo inducido con otros 15 µL de tampón Laemmli, y tras calentar la mezcla a 95ºC durante 5 min, se cargan 10µL de muestra en el gel.

Finalmente se seleccionan los cultivos que sobre-expresan proteína para su posterior purificación

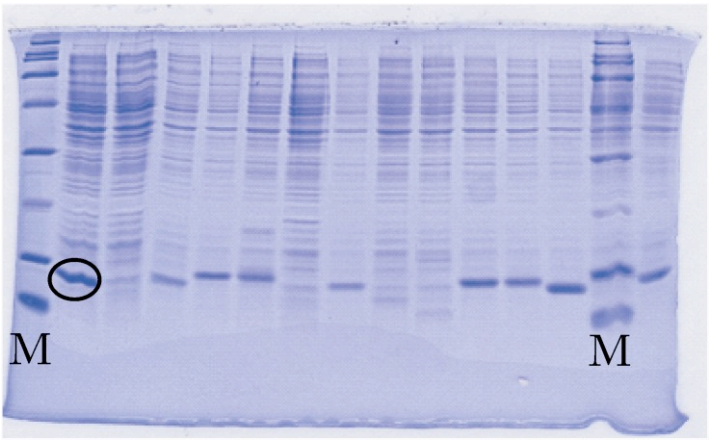

Figura 6. Ejemplo de pruebas de expresión hechas en un gel de poliacrilamida al 15%. Las calles M corresponden al marcador de peso molecular, el resto corresponden a diferentes cultivos de colonias. La banda rodeada por el círculo corresponde a la sobreexpresión de una variante de tiorredoxina.

2.2.3 Purificación de proteínas

2.2.3.1 Purificación de proteínas con cola de histidina

Con objeto de simplificar el proceso de purificación de la tiorredoxina hasta el punto de poder realizar múltiples purificaciones simultáneas de distintas variantes, optamos por agregar una cola de polihistidina de 6 histidinas a su extremo N terminal. Las colas de histidina son comúnmente usadas para facilitar la purificación de proteínas, debido a su pequeño tamaño y porque muy raramente interfieren con la estructura y función de la proteína, por lo tanto en la mayoría de los casos no se requiere su posterior eliminación.

La purificación se realiza por cromatografía de afinidad, no necesitando ninguna otra técnica cromatográfica adicional para lograr obtener una elevada pureza de muestra, que suele ser mayor del 90% en la mayoría de los casos, comprobada por SDS-PAGE. Las colas polihistidina tienen alta afinidad por el Ni^{2+} así como por otros iones metálicos; por tanto al pasar una fase móvil (parte soluble de células lisadas) a través de una fase estacionaria que contiene grupos quelantes de Ni^{2+}, este formará un complejo de coordinación con los anillos de imidazol de las histidinas. La unión de la proteína es muy inestable si estos complejos se forman con residuos de histidina dispersos por la cadena polipeptídica de la proteína, pero dicha unión se fortalece cuando la proteína dispone de varios residuos de histidina contiguos como en el caso de las colas de histidina. Esta técnica es conocida como cromatografía de afinidad de quelantes inmovilizados (IMAC).

El imidazol compite con las histidinas por los puntos de unión del Ni^{2+}, este será usado a bajas concentraciones (20 mM) para el equilibrado y lavado de la columna, ya que reducirá las posibles uniones inespecíficas de otras proteínas; también será usado a altas concentraciones (500 mM) para la elución de la proteína con cola de histidina.

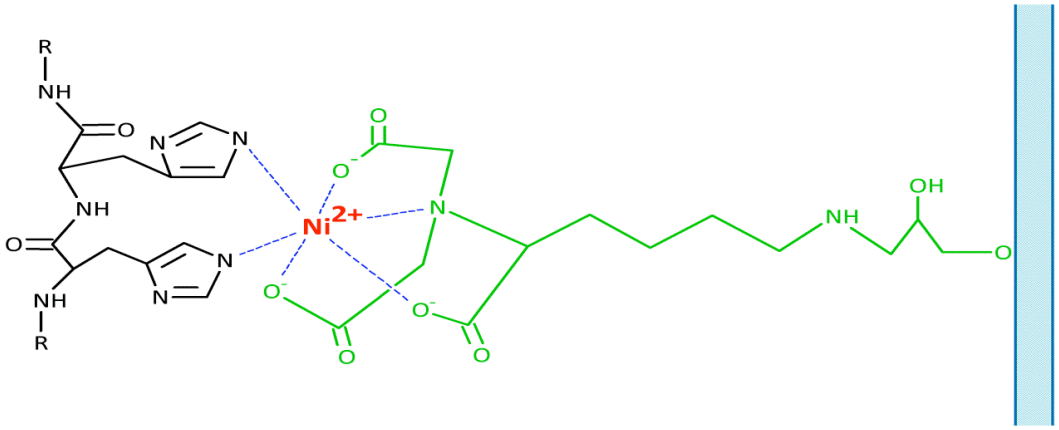

Figura 7. Esquema del complejo de coordinación del Ni^{2+} con dos histidinas de una cola de histidina (en negro) y un grupo quelante fijado a la pared de la columna de polipropileno (en verde).

El imidazol compite con las histidinas por los puntos de unión del Ni^{2+}, este será usado a bajas concentraciones (20 mM) para el equilibrado y lavado de la columna, ya que reducirá las posibles uniones inespecíficas de otras proteínas; también será usado a altas concentraciones (500 mM) para la elución de la proteína con cola de histidina.

Los tampones utilizados para la purificación además de diferentes concentraciones de imidazol llevan una alta concentración de NaCl (0.5 M) que evita interacciones electrostáticas entre proteínas y entre proteínas y ADN. El pH de 7.4 garantizará que las histidinas de la cola se encuentren desprotonadas para poder unirse al Ni^{2+}.

Para la purificación se utilizaron las columnas pre-empaquetadas His GraviTrap de GE Healthcare, la fase estacionaria es una matriz reticulada de agarosa al 6% que viene ya precargada con Ni^{2+}. En este tipo de columnas la muestra (fase móvil) pasa a través de ellas por gravedad, no requiriendo sobrepresión.

Protocolo de purificación por:

a. Cultivo de células, se inoculan con una colonia transformante o desde un cultivo preservado con glicerol 250 mL de medio de cultivo LB a 30µg/mL de kanamicina y se incuban en agitación a 37ºC.

b. Cuando el cultivo alcance la densidad óptica de 0.6 a 600nm se añade IPTG al cultivo hasta una concentración final de 0.4 mM, y se vuelve a incubar en agitación a 37ºC durante 6 a 18 horas.

c. Se recogen las células de los cultivos centrifugando a 7000 rpm durante 10 min. a 4ºC, desechando el sobrenadante.

d. Se resuspenden los pellet en 10mL de tampón de unión (20mM fosfato sódico, 500 mM NaCl y 20 mM Imidazol pH: 7.4). Con este mismo tampón equilibramos las columnas de níquel, se pasan 20mL, utilizamos las columnitas pre-empaquetadas His GraviTrap de GE Healthcare.

e. Lisamos las células por sonicación, siempre en hielo damos a cada muestra 12 pulsos de 10 segundos, con descanso entre pulso y pulso de 30 segundos.

f. Centrifugamos a 4000 rpm durante 30 min. a 4ºC en una centrifuga de mesa.

g. Filtramos el sobrenadante con filtros de 0.45 µm y decantamos el extracto crudo resultante en la columna de níquel pre-equilibrada.

h. Lavamos con 20 mL del tampón con el que se equilibró la columna.

i. Se eluye nuestra proteína en 3 mL de tampón de elución (20 mM fosfato sódico, 500 mM NaCl y 500 mM imidazol, pH: 7.4)

Tras la purificación se realizó un cambio de tampón a las muestras, que se hizo o bien mediante una diálisis o bien mediante columnas PD10 Fast Desalting (de GE Healthcare), que son columnas de exclusión molecular que usan una resina Sephadex G-25 que permite separar moléculas de bajo peso molecular de proteínas de más de 5000 Da, así las proteínas atravesarán el

volumen vacío de la columna saliendo antes que las moléculas del tampón en que estaba la proteína. Por tanto si la columna estaba pre-equilibrada con otro tampón distinto la proteína eluirá con este nuevo tampón.

2.2.3.2 Purificación de tiorredoxina y variantes sin cola de histidina

En algunos casos en el presente trabajo también se purificaron proteínas sin cola de histidina, en este caso las purificaciones constaron de 2 técnicas cromatográficas:

- Por una parte a la muestra se le realizó una cromatografía de exclusión molecular que separará las moléculas por tamaño, utilizamos como fase estacionaria Sephacryl® S-100 High Resolution de Pharmacia, que es una matriz de micropartículas con forma de esfera formadas por polímeros reticulados de poliacrilamida-dextrano, estas micropartículas tienen poros de diferentes tamaños. Las moléculas saldrán de la columna en orden decreciente al tamaño.
- La segunda es una cromatografía de intercambio iónico, en ella las moléculas son separadas en función de su carga neta. La tiorredoxina silvestre así como la mayoría de las variantes de ella estudiadas tienen carga negativa al pH del tampón de purificación 8.3. En esta cromatografía la fase estacionaria, Fractogel EMD DEAE (M), contiene grupos con carga positiva de DiEtilAminoEtilo (DEAE) que interaccionaran con los iones de carga negativa que se encuentren en la fase móvil, como nuestra proteína o el ADN que pueda estar contaminando las muestras. Las elución se realiza pasando por la columna un gradiente de concentración creciente de NaCl, así el ión Cl^- se irá uniendo a las cargas de los grupos DEAE desplazando a las moléculas que estén interaccionando con ellos a medida que aumenta

su concentración, de forma que las moléculas que estaban pegadas irán eluyendo progresivamente en función de su carga neta.

El protocolo de purificación seguido fue el siguiente:

a. Cultivo de células, se inoculan con una colonia transformante o desde un cultivo preservado con glicerol 2 litros de medio de cultivo LB a 30µg/mL de kanamicina y se incuban en agitación a 37ºC.

b. Cuando el cultivo alcance la densidad óptica de 0.6 a 600nm se añade IPTG al cultivo hasta una concentración final de 0.4 mM, y se vuelve a incubar en agitación a 37ºC durante 6 a 18 horas.

c. Se recogen las células de los cultivos centrifugando a 7000 rpm durante 10 min. a 4ºC, desechando el sobrenadante.

d. Se resuspenden los pellet en 40-50 mL de tampón T.E. (30mM Tris-Cl, 1 mM EDTA y 3 mM Azida sódica pH: 8.3). Con este mismo tampón filtrado y desgasificado equilibramos la columna de exclusión molecular, pasando 2 volúmenes de la columna, que con 1 m de largo con un radio de 2.5 cm, tiene un volumen aproximado de 2 L; por tanto habremos de pasar 4 L a un flujo constante de 5 mL/min para equilibrarla.

e. Se lisan las células con la French-press o bien por sonicación en hielo (10 pulsos de 45 segundos).

f. Se centrifugan las muestras lisadas en una ultracentrífuga durante 15 min a 20000 rpm a 4ºC.

g. Se lleva a cabo un procedimiento de precipitación de ácidos nucleicos del sobrenadante, para ello:

I. Se prepara un cuarto del volumen del sobrenadante de Sulfato de estreptomicina al 10%, esta disolución se transfiere a un embudo de decantación para que con un goteo lento vaya decantando sobre

el sobrenadante que estará en constante agitación, este proceso se llevará a cabo a 4ºC.

II. Tras 8-12 horas de agitación a 4ºC la muestra se centrifuga en una ultracentrifugadora a 30000 rpm durante 30 min.

Este procedimiento elimina parcialmente los ácidos nucleicos presentes en la muestra, que precipitarán junto con el sulfato de estreptomicina.

h. Filtramos el sobrenadante con un filtro de 0.45 μm y cargamos la muestra en la columna mediante la bomba del FPLC a 5 mL/min. La columna por su tamaño acepta un volumen de carga de hasta 120 mL de muestra.

i. Una vez cargada la muestra el FPLC sigue pasando tampón T.E. a 5 mL/min y registrando en todo momento la absorbancia a 280 nm, la tiorredoxina y sus variantes por su tamaño comienzan a salir de la columna aproximadamente a los 250 min de haber cargado la muestra (1250 mL). Un colector de fracciones se encargará de ir recogiendo fracciones de nuestra proteína de 10 mL.

j. Se comprueba la pureza de la proteína de las fracciones recogidas mediante SDS-PAGE en un gel al 15% de poliacrilamida, comprobando si hay impurezas de otras proteínas. Midiendo el espectro de absorbancia podremos saber si nuestra muestra está contaminada con ácidos nucleicos. Los ácidos nucleicos dan un máximo de absorción a 260 nm mientras que las proteínas, en concreto la tiorredoxina, lo hacen mayoritariamente a 280 nm, para considerar a la muestra libre de ácidos nucleicos la relación Abs 280_{nm}/Abs 260_{nm} deberá ser ≥ 1.6

k. Para eliminar posibles restos de ADN y otras impurezas de nuestra proteína, pasaremos las fracciones de mayor pureza por la columna de

intercambio iónico; equilibrando previamente dicha columna con 2 volúmenes de T.E. pH: 8.3 a un flujo constante de 2 mL/min. Una vez cargada la muestra en la columna pasaremos un gradiente creciente de NaCl (de 0 a 500 mM en 400 mL), el FPLC registra la absorbancia a 280 nm mientras el colector de fracciones toma una fracción cada 5 mL, como las moléculas serán separadas en función de su carga neta lo normal es obtener 2 picos principales uno de ellos el de nuestra proteína pura y otro que saldrá más tarde de ácidos nucleicos.

1. Se comprueba la pureza de las fracciones recogidas mediante SDS-PAGE y por densitometría.

Una vez obtenida la proteína dializaremos en tampón más adecuado para su caracterización.

2.3 Preparación de las muestras

2.3.1 Diálisis

El peso molecular de las variantes de tiorredoxina sin cola de histidina es aproximadamente de 11.7 KDa mientras que el de las variantes con cola de histidina es aproximadamente de 13 KDa; por tanto, se utilizó, cuando hizo falta dializar los mutantes, una membrana de diálisis con tamaño de poro de 6 a 8 KDa. Las diálisis se realizaron en agitación, a 4^{o}C, y se realizó un cambio de buffer cada 6-8 horas hasta un total de 4 cambios.

2.3.2 Medida de concentración

La medida de la concentración de las proteínas estudiadas en este trabajo se realizó siempre espectrofotométricamente, midiendo su absorbancia a 280 nm. Atendiendo a la ley de Lambert-Beer:

$$A = \varepsilon l c \qquad (1)$$

Donde A es Absorbancia, ε es el coeficiente de extinción molar de la proteína, l es la distancia que la luz atraviesa (en centímetros) y c será la concentración (normalmente Molar) de la sustancia absorbente en el medio.

El coeficiente de extinción molar de la tiorredoxina silvestre es de 14105 $cm^{-1}M^{-1}$. Se recalculó el coeficiente de extinción molar para las variantes de tiorredoxina que sufrieron un aumento o disminución de los aminoácidos aromáticos (triptófanos, fenilalanina y tirosina) con respecto a la forma silvestre, dicho cálculo se realizó mediante el uso de la herramienta web: http://www.expasy.ch/tools/protparam

2.4 Caracterización de proteínas

2.4.1 Calorimetría diferencial de barrido

2.4.1.1 Introducción:

La calorimetría diferencial de barrido (DSC, *Differential Scanning Calorimetry*) es una técnica muy potente para la caracterización energética de los cambios conformacionales inducidos por temperatura en macromoléculas biológicas tales como proteínas, ácidos nucleicos y membranas. En el caso de las proteínas, que es lo que concierne a este trabajo, los estudios de DSC sobre su desnaturalización térmica han jugado un papel fundamental en el desarrollo del actual punto de vista acerca de los factores que determinan su estabilidad. Podemos encontrar en la literatura científica diversas reseñas sobre los aspectos más relevantes de esta técnica [41-48]. Un calorímetro diferencial de barrido básicamente consiste en un sistema adiabático con dos compartimentos o células que pueden calentarse a una velocidad constante (velocidad de barrido, v) en un rango de temperaturas determinado; en un experimento típico, como los que hemos realizado en este trabajo, una de las células contiene una disolución de proteína, y la otra el solvente puro; de tal

forma que se puede obtener la capacidad calorífica de la proteína en función de la temperatura. El perfil típico de DSC, comúnmente conocido como termograma, normalmente exhibe un pico o transición atribuido a la absorción de calor asociada a la desnaturalización de la proteína, que es un proceso endotérmico, mientras que a ambos lados de la transición se encuentran las capacidades caloríficas del estado nativo y del desnaturalizado.

Como veremos más adelante un posterior análisis de esta dependencia de la capacidad calorífica con la temperatura nos podrá dar una detallada información termodinámica o cinética del proceso de desnaturalización.

En particular de los experimentos de DSC que puedan ser analizados según la termodinámica de equilibrio se puede obtener información de:

- La capacidad calorífica parcial absoluta de una macromolécula.
- El conjunto de parámetros termodinámicos asociados a la transición inducida por temperatura (cambio de entalpía (ΔH), cambio de entropía (ΔS) y cambio en la capacidad calorífica a presión constante (ΔC_p))
- La función de partición y, consecuentemente la población de estados intermedios así como sus parámetros termodinámicos.

2.4.1.2 Instrumentación

En el presente trabajo se utilizó un calorímetro VP-Capillary DSC [49]. Como se acaba de mencionar, un calorímetro diferencial de barrido consiste básicamente en dos células gemelas, una para muestra y otra para referencia. Estas se encuentran en contacto íntimo con una serie de resistencias eléctricas cuyo objeto será el de aumentar su temperatura, llamaremos a estas resistencias calentadores principales y auxiliares, y serán independientes para cada una de las células. Conectando ambas células se encuentra un sensor de cristal ($Bi_2 Te_3$), que se encargará de registrar cualquier mínima diferencia de

temperatura entre ellas. Todo esto se encuentra en el interior de una coraza adiabática que está en contacto con un mecanismo Peltier, una termopila se encargará de registrar la diferencia de temperatura entre las células y la coraza. A las células se accede por dos conductos que atraviesan la coraza adiabática (ver Figura 8).

Al comenzar el barrido calorimétrico se suministra a los calentadores principales la misma intensidad de corriente (valor que es fijado por la velocidad de barrido, v, seleccionada). Una vez que empieza el barrido el calorimétrico, el sensor de cristal, empieza a detectar la diferencia de temperatura entre las dos células $(\mathbf{\Delta T_1})$, esta diferencia de temperatura se traduce en una señal eléctrica proporcional a dicha diferencia, que resulta en el suministro de una potencia extra que activará los calentadores auxiliares de la célula de menor temperatura, con objeto de igualar las temperaturas de ambas células, es decir de anular la señal $\mathbf{\Delta T_1}$. La diferencia de potencia que habrá que suministrar, en nuestro caso, a la célula de muestra para mantener $\mathbf{\Delta T_1}$=0 será registrada en todo momento en función de la temperatura. Esta potencia eléctrica diferencial (en mV) será directamente proporcional a la diferencia de capacidad calorífica de ambas células, por tanto, con un conveniente calibrado, recogeremos directamente la señal en μCal/K frente a Temperatura, constituyendo esta la magnitud fundamental de medida del instrumento.

Al mismo tiempo la termopila detecta otra diferencia de temperatura $\mathbf{\Delta T_2}$, esta vez entre las células y la coraza generando una segunda señal eléctrica, que dará paso a la activación del mecanismo Peltier unido a la coraza para calentar o enfriar ésta hasta anular la señal $\mathbf{\Delta T_2}$, asegurando de esta manera la adiabaticidad del proceso, que se consigue aproximando a cero el intercambio de calor entre las células y la coraza. Se pueden registrar barridos tanto calentando como enfriando, aunque, en el caso del enfriado este se hace

de forma no abiabática, ya que los calentadores asociados a las células solo pueden calentar, por tanto lo único que entrará en juego será el mecanismo Peltier que enfriará la coraza, el enfriamiento de las células se conseguirá por su intercambio de calor con el exterior.

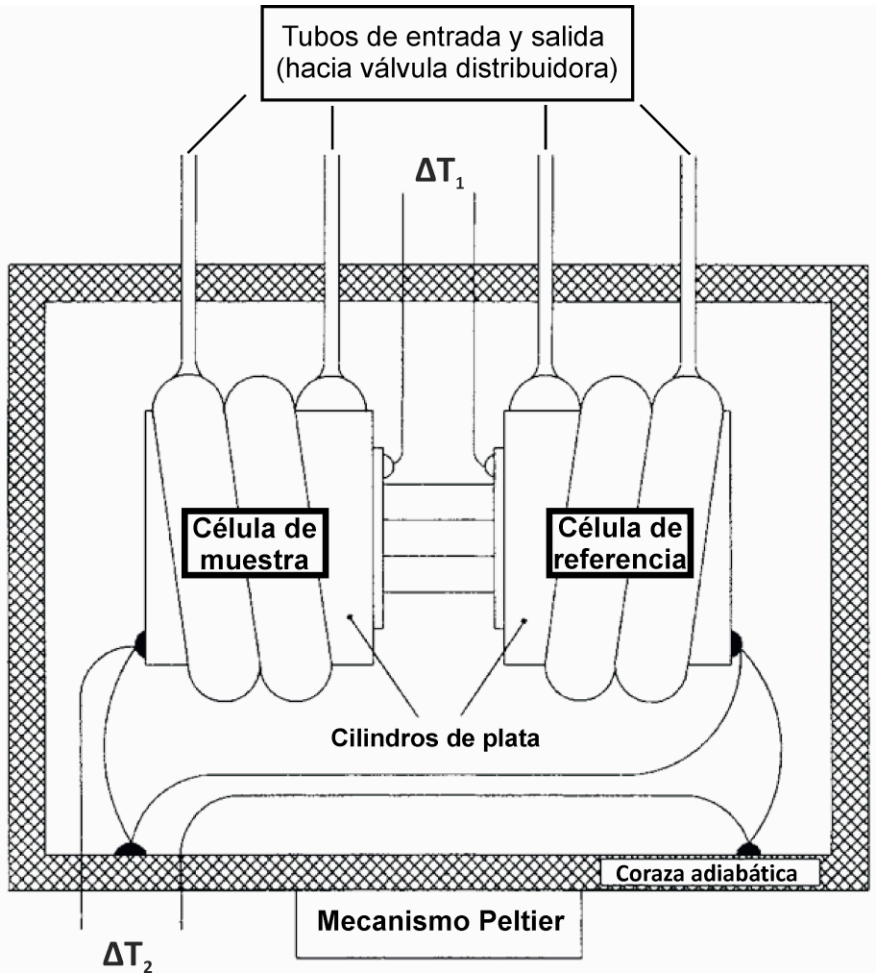

Figura 8. Esquema del núcleo calorimétrico del calorímetro capilar VP-DSC [49]. Las células se disponen en forma de hélice en torno a los cilindros de plata que contienen los calentadores. Uniendo las dos células se encuentran los 64 empalmes de Bi_2 Te_3 que conforman el sensor de cristal y que registran la diferencia de temperatura entre ambas (ΔT_1). Como se puede observar la coraza adiabática, que está unida a un mecanismo Peltier, se encuentra rodeando las células calorimétricas y en contacto con estas, y con la coraza hay una termopila que se encargará de registrar la diferencia de temperaturas (ΔT_2). Figura modificada de [49]

Estos dos circuitos de retroalimentación $\mathbf{\Delta T_1}$ y $\mathbf{\Delta T_2}$ estarán pues funcionando durante todo el proceso operativo del instrumento con objeto de mantener las dos células a la misma temperatura y la coraza a la misma temperatura que las células.

El calorímetro utilizado, VP-Capillary DSC [49], (ver Figura 9), está basado en el diseño del VP-DSC [50], aunque presenta sustanciales diferencias respecto a este:

- La principal ventaja de este calorímetro es su automatización, está equipado con un brazo robótico (ver Figura 9) que irá tomando muestras, mediante una jeringuilla Hamilton, de una placa de 96 pocillos e introduciéndolas en el calorímetro con la ayuda de una válvula distribuidora. El brazo robótico también se encargará de expulsar las muestras estudiadas (por desplazamiento) y de lavar las células antes de cargar nuevas. La automatización es fundamental para el estudio de bibliotecas de mutantes, donde las variantes a estudiar suelen ser numerosas, el hecho de poder programar secuencias de experimentos que el calorímetro pueda realizar de forma automática durante días permite aumentar el número de variantes que pueden ser estudiadas. Las placas de muestras se encuentran en una cámara refrigeradas a 4ºC.
- Conectado a este sistema de válvulas se encuentra una entrada de nitrógeno que será la que suministre sobrepresión a las células.
- Diferencias en el bloque calorimétrico:
    - Las células calorimétricas son de Tántalo, que es un material extremadamente resistente a la corrosión. Estas tienen forma de tubo con 1.5 mm de diámetro interno y se disponen en forma de hélice alrededor de un cilindro de plata, que lleva asociados los calentadores principales y auxiliares (ver Figura 8). Una alta relación superficie/volumen de las células asegura que el

equilibrado térmico de la muestra ocurra por contacto con la superficie del metal de forma homogénea evitando así la generación de corrientes de convección.

  - El volumen efectivo de cada célula es de aproximadamente 0.125 mL
  - Dos tubos de 0.8 mm, uno de entrada de muestra y otro de salida, atraviesan la coraza adiabática y comunican cada célula con la válvula distribuidora. El uso de células capilares con tubos separados de entrada y de salida facilita la limpieza y disminuye el riesgo de cargar burbuja, que altera por completo la señal cuando está presente.

- El rango de temperaturas operativo va desde -10 a 130 ºC, pudiéndose seleccionar velocidades de barrido entre 0 y 250 ºC/hora.

Figura 9. Imagen de un calorímetro VP-Capillary DSC. El brazo mecánico se mueve a través de la barra desplazadora permitiendo el proceso de automatización. Las placas que contienen las muestras se sitúan en el interior de una cámara refrigerada a 4ºC. Imagen de http://www.genengnews.com

2.4.1.3 Experimento calorimétrico

2.4.1.3.1 Preparación de las muestras

Un calorímetro es un instrumento de medida muy preciso, y dado que de sus medidas se van a inferir parámetros termodinámicos tales como el cambio de entalpía del proceso de desplegamiento de una proteína, las muestras de proteínas han de prepararse de una forma muy rigurosa, de manera que se garantice una adecuada pureza e integridad de la muestra así como una cuantificación muy precisa de su concentración, de lo contrario no podremos determinar con exactitud parámetros termodinámicos. En consecuencia para la preparación de las muestras se ha procedido de la siguiente manera:

- **a.** Tras la purificación se ha comprobado la pureza de las proteínas mediante SDS-PAGE., desechando aquellas muestras que tuvieran menos del 95% de pureza.
- **b.** Las muestras de proteínas se han dializado frente al tampón correspondiente como se ha descrito en el apartado (diálisis).
- **c.** Tras la diálisis dichas muestras se han centrifugado a 15000 rpm durante al menos 30 min con objeto de eliminar los agregados de proteína que se hubieran podido formar y que podrían distorsionar la medida.
- **d.** La medida de la concentración de proteína se realizó como ha sido descrito anteriormente.

Tras la preparación de las muestras se cargó un volumen de 0.4 ml en los pocillos de la placa tanto para las muestras de proteínas como para sus referencias, ya que si bien el volumen efectivo de las células del calorímetro esta en torno a 0.125 ml, el calorímetro ha de tomar muestra en exceso para garantizar la no intrusión de burbujas dentro de las células calorimétricas.

Como referencias de los experimentos calorimétricos se utilizó siempre el tampón resultante del último cambio de diálisis de las muestras.

2.4.1.3.2 Línea base instrumental

Las células del calorímetro como cabe esperar no son exactamente iguales, si lo fueran bastaría un solo barrido muestra-referencia para determinar la diferencia en capacidad calorífica de ambas disoluciones; por tanto es necesario realizar barridos referencia-referencia previos, se harán varios ya que los primeros barridos no salen reproducibles, y se restará el último al barrido muestra-referencia con objeto de eliminar la contribución meramente instrumental de la señal. Dicha sustracción dará lugar, dentro de la mayor parte del intervalo de temperaturas registrado, a valores negativos de capacidad calorífica, esto se debe a que la capacidad calorífica de la referencia es mayor que la de la muestra (ver Figura 10), la capacidad calorífica del agua líquida es mayor que la de la proteína, y el contenido de agua es menor en la muestra, ya que la proteína ha desplazado al agua.

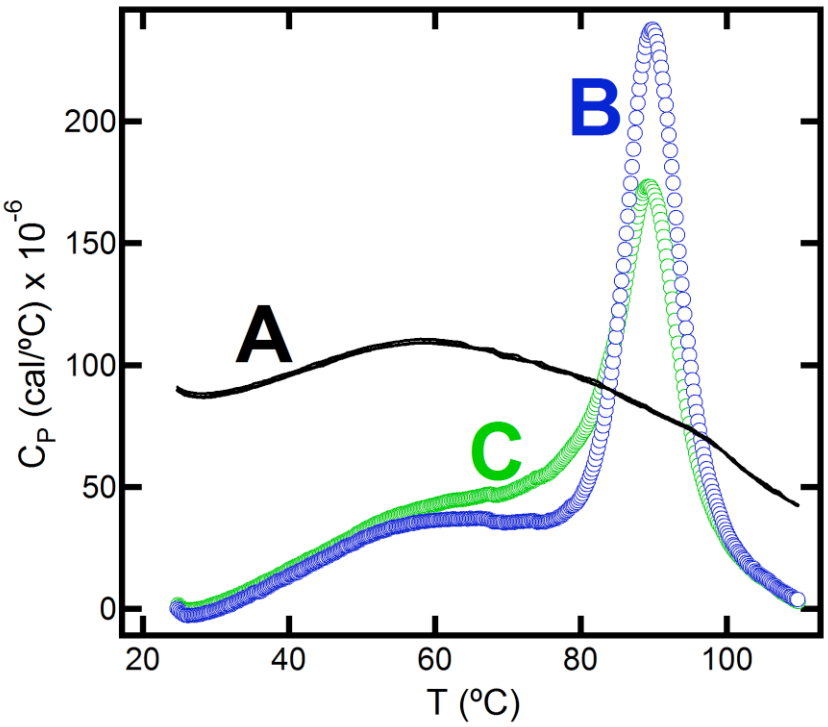

Figura 10. Gráfico con datos originales de un experimento de DSC. (A) Corresponde a 5 líneas base instrumentales solapadas; (B) es el barrido de tampón-proteína, y (C) es el segundo barrido (recalentamiento) de la misma muestra de (B) tras su enfriamiento. Los datos mostrados en este gráfico son experimentales, y se obtuvieron bajo las siguientes condiciones: la proteína del experimento es un mutante de tiorredoxina de *E. coli* a una concentración de 35.5µM en HEPES 5 mM pH: 7, siendo por tanto la referencia este mismo tampón. La velocidad de barrido del experimento fue de 1.5 K/min.

2.4.1.3.3 Barrido de muestra

Como fue comentado en el apartado anterior, antes de que comience el barrido de la primera muestra de proteína de una tanda de experimentos, se programa el registro de un número razonable de líneas base (entre 4 y 6) con objeto de conseguir un grado de reproducibilidad que esté dentro de las especificaciones del calorímetro (±2µcal/grado). Entre las subsiguientes muestras de proteína el registro de 1 o 2 líneas base se consideraron suficientes, siempre y cuando no se rompieran las condiciones de reproducibilidad correspondientes. Para asegurar dicha reproducibilidad las condiciones programadas para los barridos (temperaturas de inicio y final, velocidad de barrido, tiempo de equilibrado del calorímetro entre el principio de un barrido y en comienzo del otro) fueron siempre las mismas. El calorímetro realiza el vaciado y posterior llenado de las células siempre a una temperatura cercana a la temperatura ambiente, 25ºC, para evitar desequilibrios. Las muestras se prepararon a concentraciones de entre 0.2 y 2 mg/ml. Tras acabar el barrido de una muestra, fue siempre programado un segundo barrido a esta misma muestra, lo que comúnmente se conoce como ''reheating'' (recalentamiento) con el propósito de verificar la reversibilidad del proceso de desplegamiento de la proteína (ver Figura 10); generalmente dicha reversibilidad se expresa en términos del porcentaje de área bajo la curva correspondiente al primer barrido de muestra, que se recupera en el segundo barrido de la muestra. Este es un aspecto fundamental a la hora de analizar los datos, ya que si no existe reversibilidad no se podrá aplicar un

análisis basado en la termodinámica de equilibrio, teniendo que recurrir a otro tipo de modelos.

2.4.1.4 Análisis de datos de DSC

Una vez acabado el experimento calorimétrico, tras realizar un tratamiento previo de los datos consistente en sustraer la línea base instrumental a la de la muestra y en pasar los datos a las unidades adecuadas ($kJ \cdot K^{-1} \cdot mol^{-1}$), procedemos a su análisis con objeto de obtener la máxima información posible, ya sea termodinámica o cinética, del proceso de desnaturalización térmica. El modelo matemático a utilizar será elegido en base a una serie de criterios [48, 51, 52]:

- Reversibilidad del proceso: Si la proteína una vez desnaturalizada térmicamente es capaz o no de retornar a su estado nativo tras volver a temperatura ambiente; hecho que se puede verificar experimentalmente con el ''reheating''. En caso de existir reversibilidad estaremos ante un proceso en equilibrio pudiendo aplicar al análisis la termodinámica de equilibrio. En caso contrario se tratará de un proceso cinético que tendrá que ser descrito por ecuaciones de velocidad.
- Efecto de la concentración: Si la temperatura a la que se da el máximo de la transición (T_m) es independiente de la concentración de proteína, o si por el contrario aumenta a medida que disminuye la concentración, lo cual resultaría indicativo de procesos de oligomerización, de cooperatividad intermolecular en el proceso de desnaturalización.
- Efecto de la velocidad de barrido: Si la transición no se ve afectada por la velocidad de barrido a la que se haga el experimento, o si por el contrario la transición sufre un desplazamiento cuando el experimento

lo hacemos a varias velocidades de barrido, en cuyo caso tendremos una distorsión cinética del proceso de desplegamiento.

- Unión a ligandos: Si la proteína une un ligando tendrá efecto estabilizador, es decir aumentará la T_m de la transición.

Una vez dispongamos de toda la información experimental necesaria podremos elegir el modelo que mejor y de una forma más simple explique los resultados.

2.4.1.4.1 Análisis de los datos según la termodinámica de equilibrio.

Las proteínas estudiadas bajo las condiciones experimentales realizadas en este trabajo se han podido analizar con un modelo sujeto a las leyes de la termodinámica de equilibrio, en el cual suponemos que en todo momento existe equilibrio entre los diferentes estados poblados de estas proteínas. Dentro de estos modelos nuestros datos experimentales pudieron explicarse satisfactoriamente por el más simple de ellos, el cual nos disponemos a exponer brevemente a continuación.

2.4.1.4.1.1 Modelo de equilibrio de dos estados.

Es el más simple de los modelos basados en la termodinámica de equilibrio. Este modelo supone que en el proceso de desnaturalización de la proteína solamente existen dos estados de la misma que están significativamente representados. La proteína puede estar en estado nativo (N) o desnaturalizado (D), y la proporción de ambos estados a una temperatura dada vendrá dado por la constante de equilibrio (K) a esa temperatura.

$$N \overset{K}{\leftrightarrow} D \tag{2}$$

$$K = \frac{[D]}{[N]} \tag{3}$$

El cambio de un determinado parámetro termodinámico debido a la desnaturalización térmica de la proteína ($\Delta_N^D J$) será definido como la diferencia de valores del estado desnaturalizado y el estado nativo para dicho parámetro.

$$\Delta_N^D J = J(D) - J(N) \tag{4}$$

Esta definición supone tomar el estado nativo como estado de referencia, por tanto podremos expresar la capacidad calorífica de exceso (C_p^{ex}), hablamos de exceso respecto al estado de referencia, así como la entalpía de exceso $\langle \Delta H \rangle$ de la siguiente manera:

$$\langle \Delta H \rangle = x_D \cdot \Delta_N^D H = \frac{K \cdot \Delta_N^D H}{1+K} \tag{5}$$

$$C_p^{ex} = \frac{\partial \langle \Delta H \rangle}{\partial T} = \frac{(\Delta_N^D H)^2}{RT^2} \cdot \frac{K}{(1+K)^2} + x_D \cdot \Delta_N^D C_p \tag{6}$$

Donde X_D [= K /(1+K)] es la fracción de proteína presente en estado desnaturalizado (D). Para expresar la dependencia de la temperatura de K y de ΔH se hizo uso de las ecuaciones de Van't Hoff (7) y de Kirchoff (8).

$$\frac{\partial \ln K}{\partial T} = \frac{\Delta_N^D H}{RT^2} \tag{7}$$

$$\frac{\partial \langle \Delta H \rangle}{\partial T} = \Delta_N^D C_p \tag{8}$$

La temperatura de desnaturalización (T_m) se define como la temperatura a la cual $X_D = X_N$, por consiguiente K=1 y $\Delta_N^D G = 0$; como las transiciones predichas por la ecuación (6) son prácticamente simétricas el valor T_m estará próximo al de la temperatura correspondiente al máximo de capacidad calorífica.

Como vemos en la ecuación (6) la función capacidad calorífica de exceso consta de dos términos, el primero (término de la izquierda) refleja la transición, es decir el cambio inducido por la temperatura en el equilibrio de desnaturalización, teniendo forma de pico, mientras que el segundo término

representa la capacidad calorífica promedio de la proteína lo que se conoce como la línea base química, de forma sigmoidal (ver Figura 11).

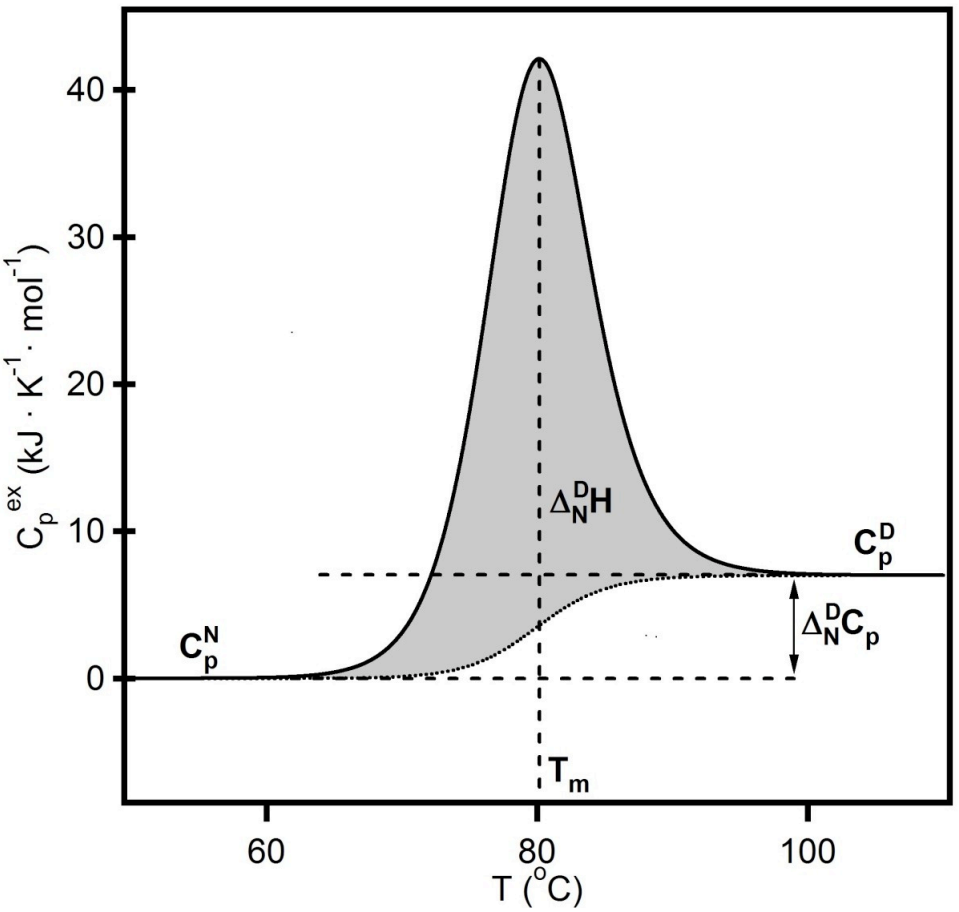

Figura 11. (Curva de trazo continuo) Curva simulada de la dependencia de la capacidad calorífica de exceso con la temperatura de una muestra con proteína que sigue un modelo de equilibrio de dos estados en su desnaturalización térmica. Donde C_p^N y C_p^D , que se ubican al principio y al final de la curva, representan las capacidades caloríficas de los estados nativo y desplegado. (Curva de trazo punteado) Línea base química. El área sombreada corresponde a la entalpía calorimétrica $\Delta_N^D H$ también expresada como entalpía de desnaturalización $\Delta_N^D H\,(T_m)$. $\Delta_N^D C_p$ es el incremento de capacidad calorífica resultante del proceso de desnaturalización térmica y T_m es conocida como la temperatura de desnaturalización. La simulación ha sido hecha utilizando los siguientes parámetros: $\Delta_N^D C_p$= 7 kJ·K^{-1}·mol^{-1}, $\Delta_N^D H$ = 400 kJ·mol^{-1} y T_m = 80 ^{o}C.

El área encerrada entre la transición y la línea base química (los dos términos de la ecuación (6)) (ver Figura 11) corresponde al cambio de entalpía total del

proceso de desnaturalización térmica, que se conoce como *entalpía calorimétrica* ($\Delta_N^D H$). Ésta depende de la temperatura y del ($\Delta_N^D C_p$) como describe la ecuación de Kirchoff (8). Es fácilmente demostrable que no se comete mucho error si se asigna la entalpía calorimétrica a la temperatura de desnaturalización [$\Delta_N^D H(T_m)$]. Dado que:

$$K=e^{(-\Delta_N^D G/RT)} \tag{9}$$

Cuando $T=T_m$, $K=1$ y por tanto el cambio de energía libre a la temperatura de desnaturalización es 0 [$\Delta_N^D G(T_m)=0$], por tanto la entropía a esta temperatura vendrá dada por la siguiente expresión:

$$\Delta_N^D S(T_m)=\frac{\Delta_N^D H(T_m)}{T_m} \tag{10}$$

Como la capacidad calorífica de la desnaturalización [$\Delta_N^D C_p$] puede obtenerse directamente del perfil calorimétrico como se puede ver en la Figura 11, podemos obtener los valores de $\Delta_N^D G$, $\Delta_N^D H$ y de $\Delta_N^D S$ a cualquier temperatura usando las ecuaciones termodinámicas convencionales:

$$\Delta_N^D H(T)=\Delta_N^D H(T_m)+\int_{T_m}^{T}\Delta_N^D C_p \cdot dT \tag{11}$$

$$\Delta_N^D S(T)=\frac{\Delta_N^D H(T_m)}{T_m}+\int_{T_m}^{T}\frac{\Delta_N^D C_p}{T} \cdot dT \tag{12}$$

$$\Delta_N^D G(T)=\Delta_N^D H(T)-T \cdot \Delta_N^D S(T) \tag{13}$$

Por supuesto todo lo descrito anteriormente asume que el proceso de desnaturalización estudiado puede ser explicado por el modelo de equilibrio de dos estados. La entalpía calorimétrica de desnaturalización [$\Delta_N^D H(T_m)$] determina tanto al área encerrada entre la transición y la línea base química, como forma (anchura) de la transición. Mientras que la entalpía que puede ser calculada únicamente a través de la anchura de la transición es conocida como entalpía aparente o de Van't Hoff [ΔH^{vH}]. Podemos obtener la siguiente

expresión para calcular [ΔH^{vH}] a partir del primer término de la ecuación (6) si consideramos $T=T_m$, y por tanto K=1:

$$\Delta H^{vH}=4RT^2 \cdot \left[\frac{\Delta C_p(T_m)}{\Delta_N^D H(T_m)}\right] \quad (14)$$

Donde [$\Delta C_p(T_m)$] será el cambio de capacidad calorífica a la temperatura de desnaturalización que será medida desde la línea base química, (ver Figura 11); el cociente de esta y la entalpía calorimétrica [$\Delta C_p(T_m) / \Delta_N^D H(T_m)$] no va a depender de la cantidad de proteína presente en el experimento, si no de la forma de la transición.

Como cabe esperar no pueden existir dos variaciones de entalpía para un mismo proceso, la ecuación (14) es inferida de la ecuación (6), la cual está sustentada bajo el supuesto de encontrarnos ante un equilibrio de dos estados, por tanto si el modelo es de equilibrio de dos estados ambas entalpías deben de coincidir (dentro del error experimental), así el siguiente cociente, r, debería de ser igual a 1.

$$r=\frac{\Delta_N^D H(T_m)}{\Delta H^{vH}} \quad (15)$$

Si por el contrario ambas entalpías difieren de forma significativa, $r \neq 1$, podemos descartar que el proceso de desnaturalización que estamos estudiando pueda explicarse por este modelo. Este cociente es conocido como el ''test de dos estados'', una de las posibles causas de una desviación de la unidad de este cociente podría ser la existencia de otros estados que, aparte de los considerados para este modelo (N y D), estuvieran significativamente poblados en el proceso de desplegamiento (estados intermedios), la transición en este caso sería más ancha que la predicha por el modelo de equilibrio de dos estados, y por tanto $r > 1$. Otra de las posibles causas tendría que ver con posibles efectos de cooperatividad intermolecular (oligomerización) que puedan darse en el proceso de desplegamiento que supondrían un $r < 1$,

ΔH^{vH} sería mayor de lo esperado, ya que esta entalpía está calculada por mol de monómero obviando la existencia de un posible efecto multimérico.

La estabilidad de una proteína se define como la diferencia entre las energías libres de Gibbs de sus estados desnaturalizado y nativo $\Delta_N^D G$. Pero la estabilidad de la proteína dependerá de la temperatura, esta dependencia, $\Delta_N^D G$ frente a T, se conoce como curva de estabilidad (Figura 12) [53, 54]. Si el proceso de desnaturalización seguido por la proteína corresponde al previamente descrito equilibrio de dos estados podremos calcular la curva de estabilidad a partir de las ecuaciones (11)-(13), y si tomamos $\Delta_N^D C_p$ como un valor constante (en realidad tiene cierta dependencia con la temperatura, pero esta se puede ignorar sin incurrir en un error significativo) podríamos expresar las ecuaciones (11)-(13) de la siguiente manera:

$$\Delta_N^D H(T)=\Delta_N^D H(T_m)+\Delta_N^D C_p \cdot(T-T_m) \tag{16}$$

$$\Delta_N^D S(T)=\frac{\Delta_N^D H(T_m)}{T_m}+\Delta_N^D C_p \cdot \ln\left(\frac{T}{T_m}\right) \tag{17}$$

$$\Delta_N^D G(T)=\Delta_N^D H(T_m)\cdot\left[1-\frac{T}{T_m}\right]+\Delta_N^D C_p \cdot\left[T-T_m-T\cdot\ln\left(\frac{T}{T_m}\right)\right] \tag{18}$$

Las propiedades y consideraciones más importantes de la curva de estabilidad predicha por la ecuación (18) (ver Figura 12.) están descritas en los siguientes trabajos [53, 54]. A continuación hacemos una breve descripción de ellas. La curvatura viene dada por $[\partial^2\Delta_N^D G/\partial T^2= -\Delta_N^D C_P/T]$ y siempre será convexa ya que ΔC_P siempre es positivo, por tanto tendrá siempre valor negativo. La pendiente de la curva es $-\partial\Delta_N^D G/\partial T = \Delta_N^D S$ y muestra un máximo a la temperatura en la que $\Delta S=0$, temperatura T_S; la temperatura T_H, a la que $\Delta H=0$, es ligeramente más baja que T_S. La proteína nativa es estable en el rango de temperaturas en el que $\Delta G>0$, la curva de estabilidad tiene dos abscisas en el origen (donde $\Delta G=0$), la de más alta temperatura corresponderá con la temperatura de desnaturalización ya descrita (T_m), que fue usada para

el cálculo de la propia curva de estabilidad; la de más baja temperatura, resultado de la extrapolación de la curva, correspondería a la temperatura de desnaturalización por frío (T^*_m), probada experimentalmente en [55], que ocurre normalmente a temperaturas inferiores a 0 ºC y que parece ser una propiedad común a todas las proteínas globulares.

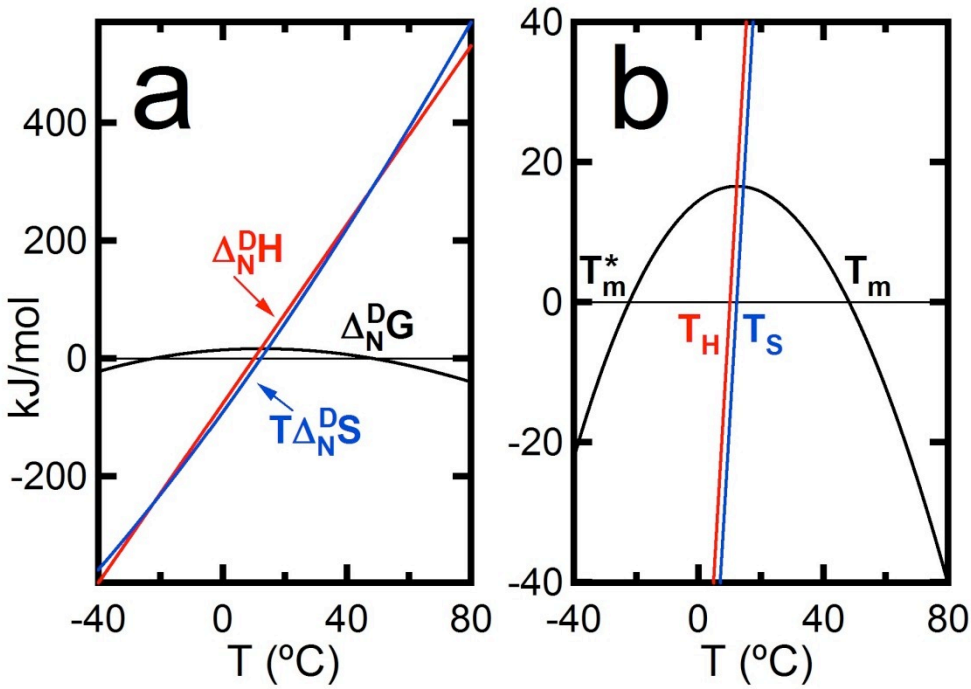

Figura 12. **(a)** Efecto de la temperatura en los cambios de entalpía, entropía y energía libre de Gibbs para la desnaturalización de una proteína hipotética. **(b)** Ampliación de la gráfica (a) para mostrar las principales características de la curva de estabilidad de la proteína (perfil de $\Delta_N^D G$ frente a temperatura), las temperaturas de desnaturalización caliente y fría, T_m y T^*_m respectivamente, y las temperaturas T_H y T_S a las cuales $\Delta_N^D H$ y $\Delta_N^D S$ son igual a cero. La simulación para calcular la dependencia de $\Delta_N^D G$, $\Delta_N^D H$ y $\Delta_N^D S$ frente a la temperatura ha sido realizada usando las ecuaciones (16)-(18) con los siguientes parámetros: T_m=48.3 ºC, $\Delta_N^D H(T_m)$=290 kJ/mol, y $\Delta_N^D C_P$=7.6 kJ·K^{-1}·mol^{-1}.

2.4.2 Dicroismo circular

El dicroísmo circular (CD, por sus siglas en inglés) es una técnica de espectroscopia de absorción que nos da información de la estructura de macromoléculas biológicas como ADN y proteínas.

El dicroísmo es el fenómeno por el cual la absorción de luz es diferente dependiendo de su dirección de polarización. En esta técnica se medirá la diferencia de cantidad de luz circularmente polarizada hacia la izquierda, sentido antihorario, (lcpi) y hacia la derecha, sentido horario, (lcpd), que es absorbida por una molécula. Para que se produzca esta diferencia de absorción dichas moléculas deben ser ópticamente activas, es decir quirales (no siendo superponibles con su imagen especular)[56].

Un haz de luz circularmente polarizada se compone a su vez de dos haces ortogonales polarizados en un plano que se encuentran fuera de fase en 90º [57, 58] Figura 13.

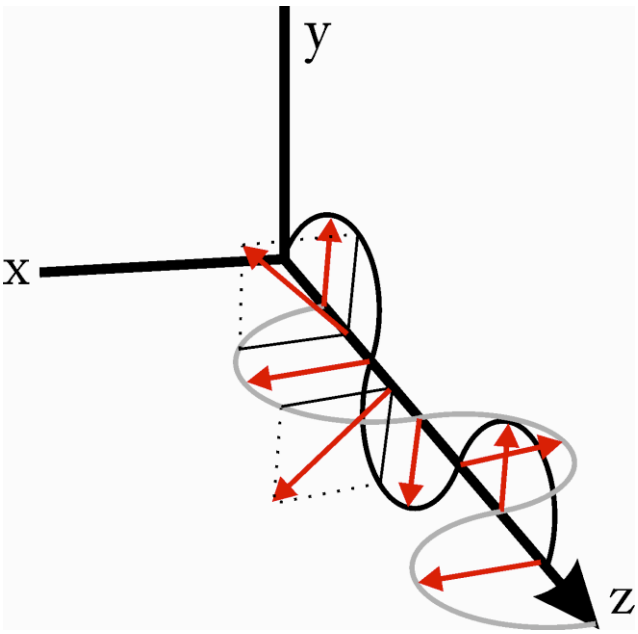

Figura 13. Haz de luz circularmente polarizada hacia la derecha. En negro y gris los haces ortogonales de luz polarizada en un plano que se encuentran fuera de fase 90º. Las flechas rojas son el vector eléctrico resultante, que como se puede observar realiza una trayectoria de rotación formando una hélice a derechas que se mueve en el sentido de propagación del haz (eje z).

Un haz de luz circularmente polarizado hacia la izquierda tendría exactamente propiedades inversas al de la Figura 13. El vector eléctrico resultante (E), que describe una trayectoria circular, se puede expresar matemáticamente de la siguiente forma:

$$E_{\pm}=E_0(\mathbf{i}\pm i\mathbf{j})e^{\left[2\pi i\left(\upsilon t-\frac{z}{\lambda}\right)\right]} \quad (19)$$

Donde + se refiere a lcpd y – a lcpi, E_0 es la amplitud de la onda, **i** y **j** son los vectores unidad en las coordenadas x e y, respectivamente, en un sistema de coordenadas dextrógiro (como el representado en la Figura 13.) en el que +Z es la dirección de propagación; *i* es $(-1)^{1/2}$ mientras que ν y λ son la frecuencia y longitud de onda del haz de luz.

El Dicroismo Circular se define como la diferencia de absorción de lcpi y lcpd por una molécula a una longitud de onda dada, así valiéndonos de la ley de Lambert-Beer podemos obtener la siguiente expresión:

$$\Delta A(\lambda)=A_i(\lambda)-A_d(\lambda)=[\varepsilon_i(\lambda)-\varepsilon_d(\lambda)]cl=\Delta\varepsilon(\lambda)cl \quad (20)$$

Donde *l* es la distancia que recorre la luz al pasar a través de la muestra, *c* es la concentración de la molécula quiral y ε el coeficiente de extinción molar de la molécula para la luz polarizada hacia el lado correspondiente a una longitud de onda. Así el CD puede expresarse en forma de Δε.

En el presente trabajo se utilizó el método de medida original de dicroísmo circular, propuesto por Thomas M. Lowry [59], que se basa en el hecho de que un haz de luz polarizada en un plano se puede descomponer en 2 haces de luz polarizados circularmente uno lcpi y otro lcpd, de la misma amplitud y frecuencia que se encuentran en fase [57, 58] (Figura 14a). Cuando dicho haz pase a través de un medio quiral, la absorción diferencial de ambos componentes producirá un cambio en la polarización de la luz, así el vector eléctrico resultante pasará de estar oscilando en un plano a trazar una elipse (Figura 14b), de forma que cuando los vectores eléctricos de los dos componentes circulares estén en el mismo sentido la suma de sus magnitudes resultará en el eje semimayor de la elipse (b en Figura 15), mientras que cuando ambos vectores estén en sentidos opuestos la resta de sus magnitudes dará el eje semimenor de dicha elipse (a en Figura 15).

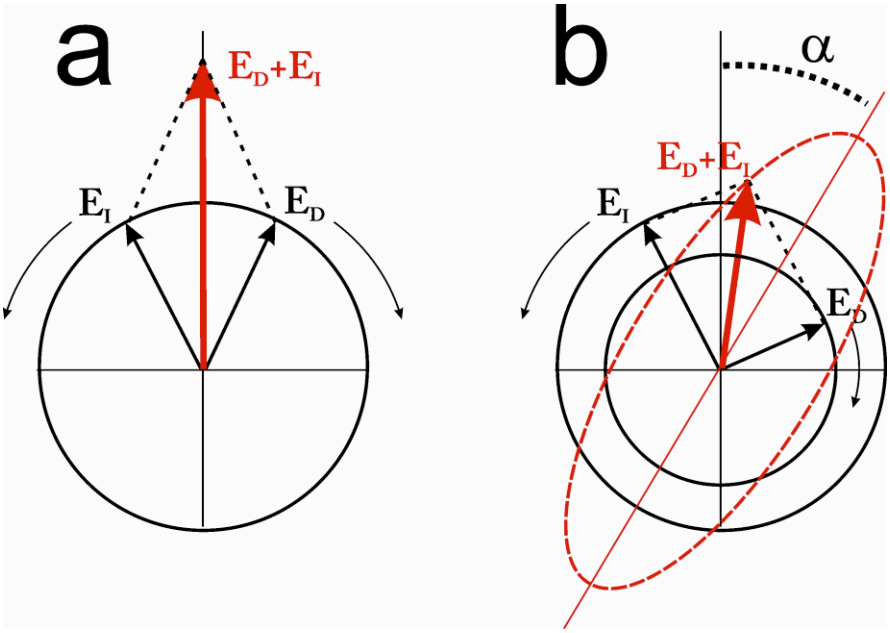

Figura 14. **a** Haz de luz polarizada en un plano, E_I y E_D son los vectores de los componentes circulares (lpcd y lcpi) que lo conforman. **b** Haz después de interaccionar con la muestra quiral, debido a la absorción diferencial de la muestra los vectores E_I y E_D han sufrido cambios distintos en sus magnitudes y se encuentran ahora fuera de fase resultando en una polarización elíptica de la luz. El ángulo α, la rotación óptica, es el formado entre el eje mayor de la elipse y el plano de polarización del haz antes de pasar por la muestra.

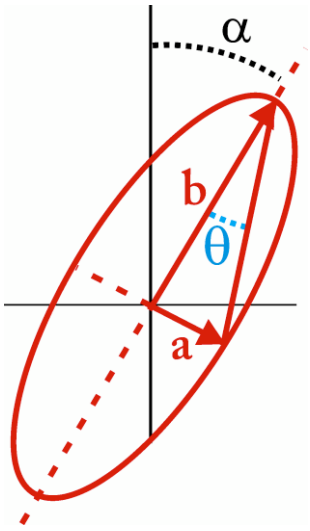

Figura 15. Elipse trazada por el vector eléctrico resultante de dos haces lcpi y lcpd de distintas intensidades. b y a son el eje semimayor y semimenor de la elipse respectivamente, el ángulo θ es la elipticidad.

El dicroismo circular puede caracterizarse por el cociente a/b, o lo que es lo mismo la tangente del ángulo θ, o elipticidad (ver Figura 15), que como es muy pequeño suele ser aproximado en radianes, así:

$$\theta(\text{rad}) \approx \tan\theta = \frac{|E_I| - |E_D|}{|E_I| + |E_D|} = \frac{e^{(-A_I/2)} - e^{(-A_D/2)}}{e^{(-A_I/2)} + e^{(-A_D/2)}} \quad (21)$$

Expandiendo las expresiones exponenciales, teniendo en cuenta el valor despreciable de ΔA comparado con la unidad y convirtiendo a grados:

$$\theta(\text{grados})_\lambda = 180 \cdot \ln 10 \cdot \frac{\Delta A_\lambda}{4\pi} = 32.98 \cdot \Delta A_\lambda \quad (22)$$

Siendo la elipticidad es directamente proporcional al dicroismo circular.

Definimos la elipticidad molar, como:

$$[\theta]_\lambda = \frac{\theta_\lambda}{10 \cdot c \cdot l} \quad (23)$$

Los experimentos de dicroísmo circular realizados en este trabajo fueron expresados en términos de elipticidad molar frente a longitud de onda, las unidades de $[\theta]_\lambda$ son (grados $\cdot$dmol^{-1} $\cdot$cm^2). En cualquier caso la conversión a la escala Δε, es muy sencilla, ya que si sustituimos la ecuación (22) en la (23) obtenemos que:

$$[\theta]_\lambda = 3298 \Delta\varepsilon(\lambda) \quad (24)$$

2.4.2.1 Dicroismo circular en proteínas

El dicroísmo circular es una técnica utilizada para obtener información estructural de proteínas, que si bien no puede resolver su estructura tridimensional como la Resonancia Magnética Nuclear o la difracción de Rayos x, si nos puede dar información importante a cerca de su estructura secundaria y terciaria, así como de cambios conformacionales asociados a interacciones entre moléculas asimétricas tales como interacciones proteína-ligando, proteína-ADN o proteína-proteína.

La estructura secundaria de una proteína puede ser caracterizada mediante un espectro de dicroísmo circular realizado a longitudes de onda inferiores a

250 nm, lo que comúnmente se conoce como ultravioleta lejano. A esas longitudes de onda absorbe el enlace peptídico, que si bien no es por sí mismo una estructura quiral, su ubicación en un entorno asimétrico debido a la estructura tridimensional de la proteína le confiere actividad óptica, así los diferentes tipos de estructura secundaria que se encuentran presentes en las proteínas dan lugar a un espectro de dicroísmo circular característico, incluidos los bucles y las zonas desplegadas (zonas no periódicas) (ver Figura 16).

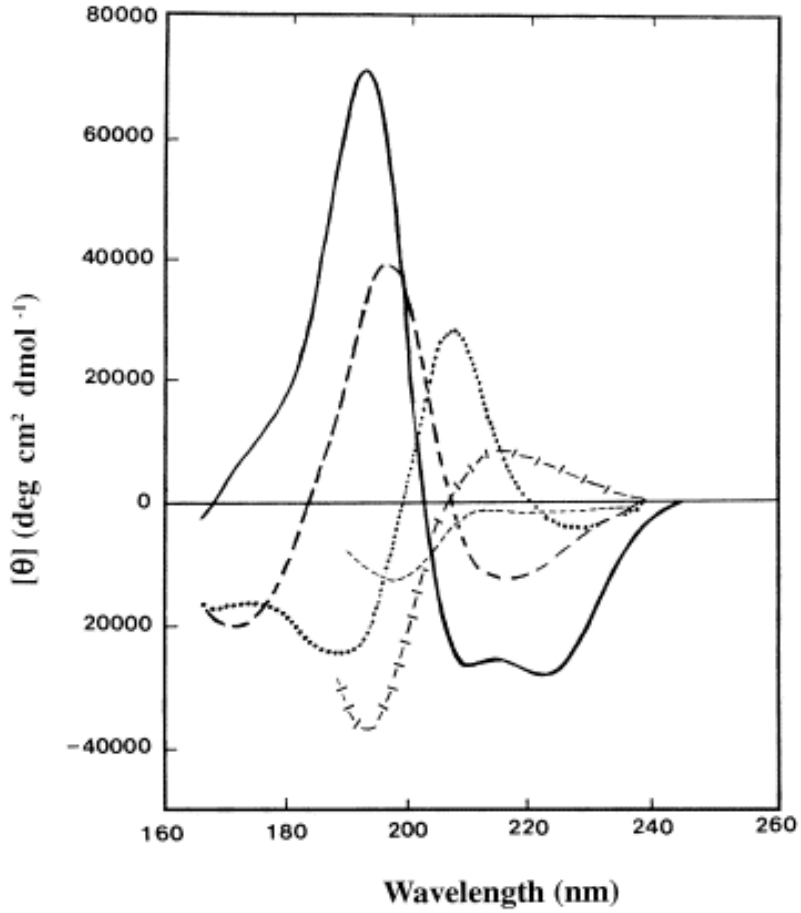

Figura 16. Espectros de DC de UV lejano asociados a varios tipos de estructura secundaria. — α-hélice, – – – lámina ß antiparalela, ····· giro ß tipo I, – 1 – 1 – hélice de poliprolina II, ---- (espectro corto) estructura irregular. Figura tomada de [60].

Cada proteína tendrá entonces un espectro de dicroísmo circular particular, de forma que sea, aproximadamente, una combinación de los espectros de

cada una de las estructuras secundarias e irregulares que posea. Además del enlace peptídico también contribuyen en cierta manera los aminoácidos aromáticos al espectro de dicroísmo circular en el UV lejano.

Los residuos aromáticos, además de otros elementos como los puentes disulfuro, también absorberán de manera diferencial la luz circularmente polarizada en el UV cercano, entre 250 y 350 nm; pero en este caso no se mezclará su contribución con la del enlace peptídico, que no absorbe a estas longitudes de onda, siendo el espectro en el UV cercano característico de la conformación nativa de la proteína, Un cambio en dicho espectro implicará cambios apreciables en la estructura nativa (terciaria) de la proteína.

2.4.2.2 Procedimiento experimental

Los experimentos de dicroísmo circular de este trabajo fueron llevados a cabo en un espectropolarímetro JASCO modelo J-715. En este trabajo se tomaron espectros de proteínas en el UV lejano entre 260 y 210 nm con objeto de determinar cambios en la estructura secundaria. A continuación, las condiciones más importantes de los experimentos realizados fueron:

- Las proteínas fueron dializadas previamente en tampón HEPES 50 mM pH:7.
- Las medidas fueron realizadas siempre a 25ºC.
- Para realizar las medidas se utilizaron cubetas de cuarzo con un paso de luz de 1mm.
- Las muestras se encontraban en un rango de concentraciones de entre 0.7-0.15 mg/mL.
- Para cada muestra se realizó el promedio de 5 espectros, de manera que se reduce significativamente el ruido del espectro resultante.

Los estudios por CD fueron realizados sobre las variantes de la biblioteca combinatorial del artículo incluido en el apartado 3.1; los espectros de estas, fueron comparados con el de la variante de partida de la biblioteca,

con objeto de determinar si se habían producido cambios importantes en su estructura secundaria.

2.4.3 Ensayos de actividad

2.4.3.1 Ensayo de actividad Reductasa

La tiorredoxina es una enzima que pertenece a la familia de las oxidorreductasas, lo implica que catalizará la transferencia de electrones desde un agente reductor a un agente oxidante. Posee un sitio activo ditiol/disulfuro, cuya forma ditiol, tiorredoxina-$(SH)_2$, se caracteriza por ser una potente reductora de puentes disulfuro de proteínas. La reducción de la forma disulfuro (tiorredoxina-S_2) es catalizada por la enzima tiorredoxina reductasa y es dependiente de NADPH [61].

La tiorredoxina también puede ser reducida por agentes distintos a los fisiológicos como es el caso del ditiotreitol (DTT) en el presente ensayo, que fue desarrollado por Arne Holmgren [62] y que seguimos en este trabajo con ligeras modificaciones.

El ensayo consiste en definitiva en medir la velocidad de reducción de la insulina por el DTT que será catalizada por la tiorredoxina (Figura 17).

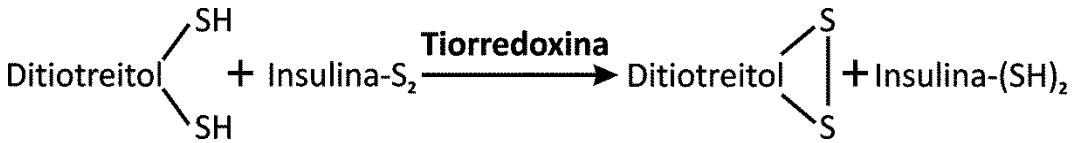

Figura 17. Esquema de la reducción de insulina, donde la tiorredoxina cataliza la reducción de la insulina por el DTT.

La insulina es una proteína compuesta por dos cadenas polipeptídicas (A y B) unidas por dos puentes disulfuro intercatenarios. Al catalizar la tiorredoxina la reacción de la Figura 17, se eliminan los puentes disulfuro, de manera que las cadenas B libres formarán, a pH neutro, agregados de alto peso molecular que constituirán un precipitado blanco insoluble, resultando en un aumento

de la turbidez, que se puede registrar espectrofotométricamente midiendo la variación de la absorbancia a 650nm con el tiempo.

La velocidad de formación de estos agregados está directamente relacionada con la actividad de la tiorredoxina, es decir con su eficiencia para catalizar la reducción de puentes disulfuro. El parámetro utilizado en el presente trabajo para la determinación cuantitativa de la actividad fue el de la velocidad máxima de formación de agregados, es decir la máxima derivada de la curva de agregación (Figura 18b); esta velocidad máxima se representa en función de la concentración de catalizador en Figura 19.

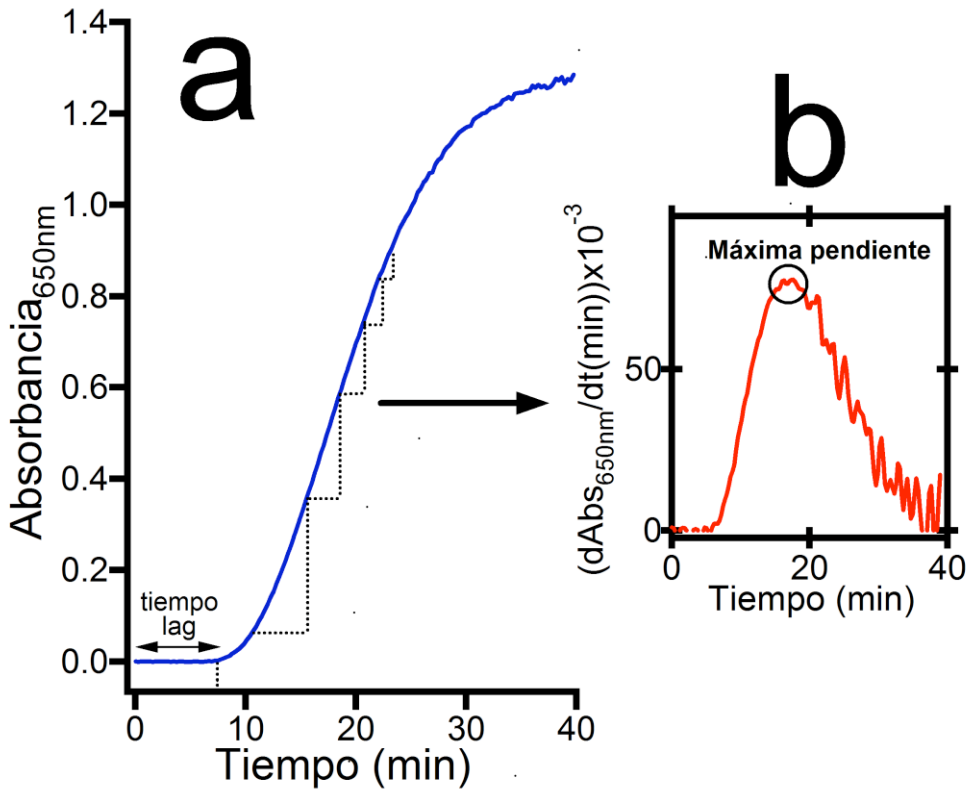

Figura 18. **a** Curva de reducción de la insulina hecha a una concentración de catalizador, tiorredoxina de *E. coli* silvestre, de 0.73 µM. El tiempo lag es tiempo transcurrido hasta que se empieza a registrar el aumento de absorbancia. Este también puede ser tomado como parámetro de medida. **b** Se muestra la gráfica de la derivada de la absorbancia respecto del tiempo, el máximo de esta gráfica se corresponde con la máxima velocidad de formación de agregados, que es el parámetro utilizado para cuantificar la actividad de la proteína catalizadora.

Para expresar la actividad de las proteínas estudiadas en este trabajo utilizamos el valor de la pendiente que resulta del ajuste de los datos a una línea recta, en la gráfica donde se representan las máximas derivadas de las curvas de reducción de la insulina frente a la concentración del catalizador (Figura 19).

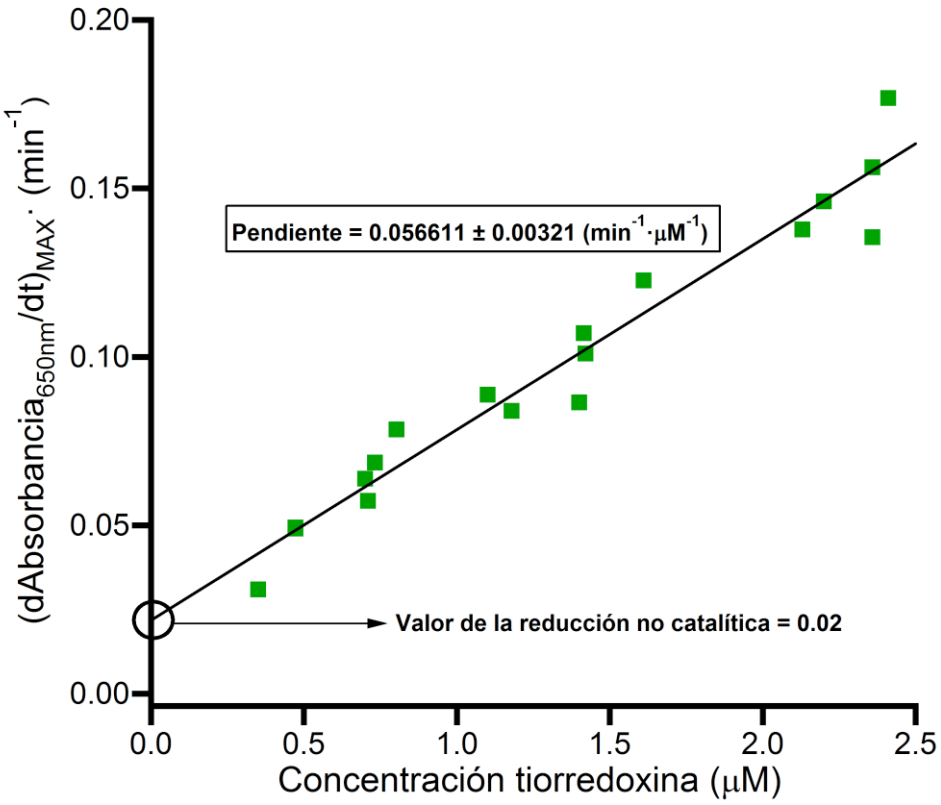

Figura 19. Recta patrón de la tiorredoxina *E. coli* silvestre para el ensayo de su actividad oxidorreductasa. Se representa la máxima derivada de la curva de reducción de la insulina frente a la concentración del catalizador. Los puntos verdes corresponden a diferentes experimentos de reducción de insulina hechos entre varios días con diferentes preparados de insulina.

2.4.3.1.1 Procedimiento experimental

2.4.3.1.1.1 Preparación de la insulina

a. Se disuelven 50 mg de insulina bovina en polvo (Sigma-Aldrich) en 4mL de tampón Tris-HCl 50 mM, pH:8.3

b. A fin de eliminar posibles agregados se realiza un salto de pH, primero se baja el pH de la disolución hasta un valor de pH de entre 2 y 3 añadiendo HCl 1M, y una vez que la disolución cambia su color blanquecino por transparente se vuelve a subir el pH de la disolución hasta 8.3 adicionando NaOH 1M.

c. Se ajusta el volumen a 5 mL con agua doblemente destilada, por lo tanto la disolución habrá de quedar a una disolución final de 10 mg/ml (1.67 mM).

d. Se hacen alícuotas y se congelan a -20ºC.

2.4.3.1.1.2 Ensayo de actividad

El ensayo de actividad se realiza a 37ºC, para la medida de la absorbancia se usó un espectrofotómetro AGILENT 8453 UV-VISIBLE conectado a un baño termostático.

a. Mezclamos los siguientes reactivos en una cubeta de plástico en el siguiente orden:

I. 784 µL de tampón de reacción (fosfato potásico 0.1 M, 2mM de EDTA, pH: 6.5) previamente termostatizado a 37ºC.

II. 50 µL del preparado de insulina descrito anteriormente.

III. 166 µL de proteína a ensayar (las proteínas se encontraban en HEPES 5 mM, pH:7)

IV. 10 µL de DTT 0.1 M.

b. A continuación se mezclan los reactivos cuidadosamente y sin demora con una pipeta pasteur. Iniciándose acto seguido la medida del espectrofotómetro a 650 mm. El proceso de medición durará 40 minutos.

Siempre que se hicieron ensayos de actividad sobre variantes de tiorredoxina, se realizaron previamente varios experimentos con la forma silvestre a fin de comprobar que los datos obtenidos caían dentro de la recta patrón (Figura

19). En los ensayos se utilizaron concentraciones finales de no más de 3 μM de proteína con el fin de trabajar siempre dentro del rango donde se mantiene la linealidad.

El valor de la reducción no catalítica puede obtenerse realizando el ensayo descrito sin añadir la enzima, de esta manera obtenemos la velocidad de reducción de los puentes disulfuro de la insulina por el DTT. También puede obtenerse por extrapolación de la recta patrón (ver Figura 19).

2.4.3.2 Ensayos de actividad de formación e isomerización de puentes disulfuro

La formación e isomerización de puentes disulfuro son dos actividades catalíticas que al igual que la actividad reductasa son realizadas por proteínas de la familia de ditiol/disulfuro oxidorreductasas, a la que pertenece superfamilia de tiorredoxina [63]. Varias enzimas de esta superfamilia, como PDI (Proteína Disulfuro Isomerasa) catalizan la formación e isomerización de puentes disulfuro entre residuos de cisteína (Figura 20), lo cual le hace estar directamente implicada en el proceso de plegamiento de muchas proteínas.

Los ensayos dirigidos a cuantificar la actividad catalizadora de formación e isomerización de puentes disulfuro se realizaron, con algunas modificaciones, como está descrito por Lundström, J. *et al* [64]. En estos ensayos se utilizó como sustrato una proteína, la ribonucleasa, que ha sido frecuentemente utilizada en estudios de plegamiento y formación de puentes disulfuro [65, 66]. Esta proteína contiene 4 puentes disulfuro en su estado nativo. En el caso del ensayo destinado a cuantificar la actividad formadora de puentes disulfuro, la ribonucleasa estaba reducida, es decir sin puentes disulfuro; y para el ensayo que cuantifica la actividad Isomerasa, se utilizó como sustrato ribonucleasa con los puentes disulfuro mal formados, estas dos formas de la ribonucleasa son inactivas. En ambos ensayos las proteínas catalizadoras se incubaron un determinado tiempo junto con las proteínas sustrato, de forma

que a mayor capacidad catalizadora mayor será la cantidad de ribonucleasa que recupere su estado nativo, y por tanto su actividad. La recuperación de dicha actividad se puede medir por la hidrólisis del 2',3' –cCMP, ver esquema Figura 21, reacción que se puede registrar con un espectrofotómetro ya que produce un incremento en la absorbancia a 288 nm [67] (ver Figura 22).

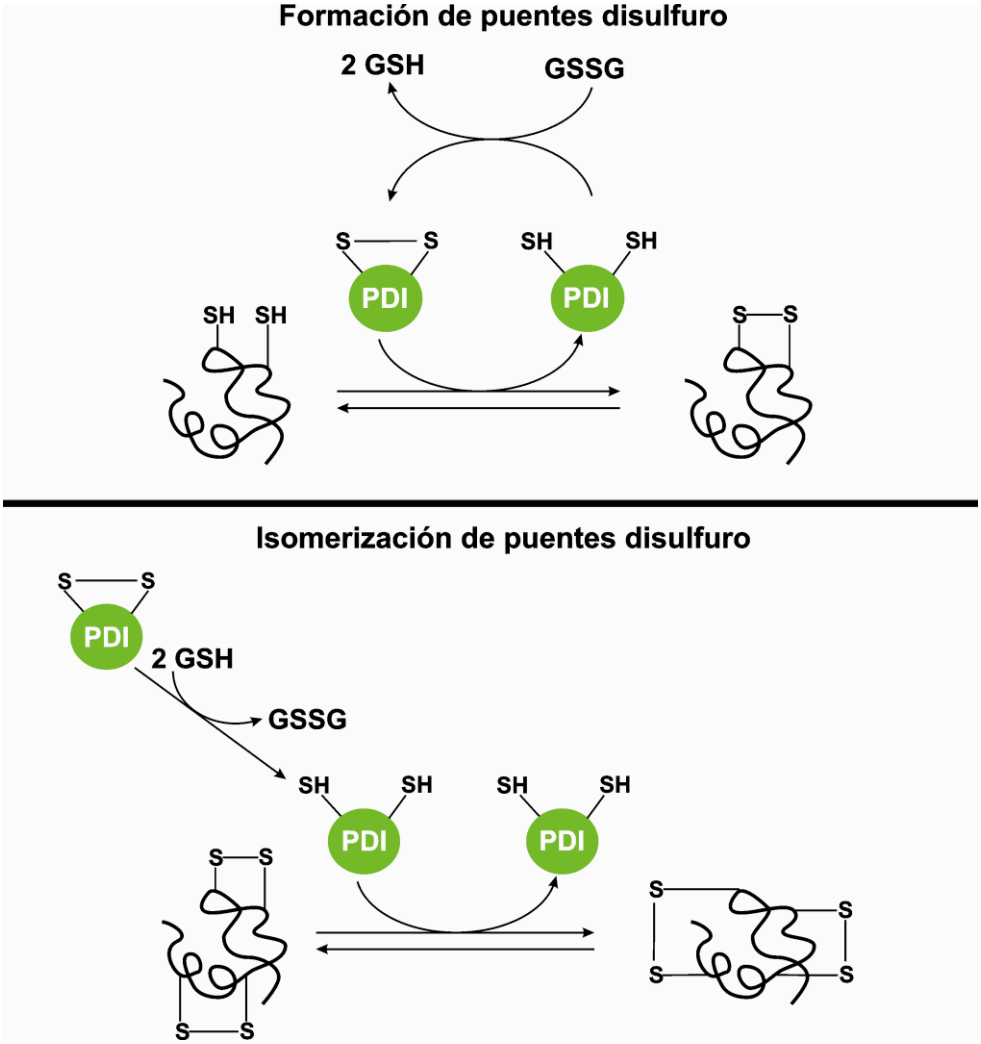

Figura 20. Catálisis de la formación e isomerización de puentes disulfuro *in vitro*. En el esquema de arriba la PDI cataliza la oxidación de dos grupos tiol de una proteína, con la consiguiente formación de un puente disulfuro utilizando glutatión oxidado como aceptor de electrones. En el esquema de abajo la PDI cataliza la isomerización de los puentes disulfuro de una proteína, la isomerización engloba ciclos de oxidación/reducción, el proceso necesita de un donador de electrones, glutatión reducido, para iniciar la reducción.

Figura 21. Esquema hidrólisis del 2′, 3′ cCMP catalizada por la ribonuclasa A. La reacción se sigue midiendo el incremento de absorbancia con el tiempo a 288 nm.

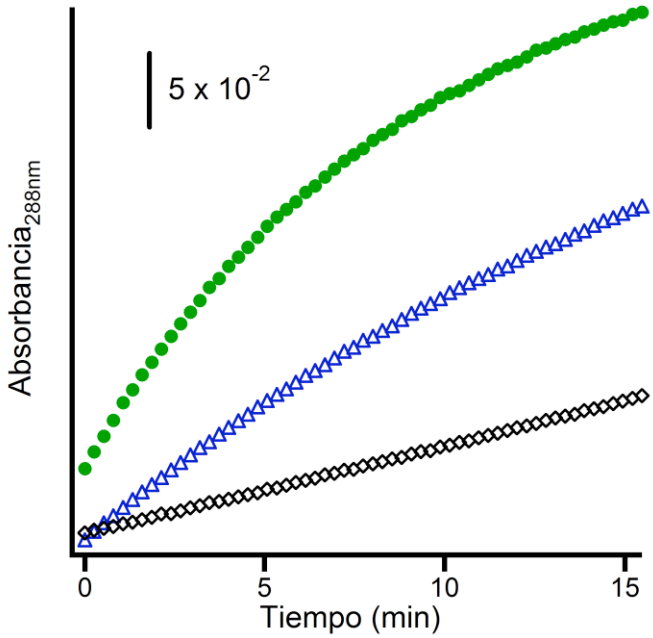

Figura 22. Cinéticas de hidrólisis del 2′, 3′ –cCMP a 25ºC catalizadas por la ribonucleasa, que fue previamente incubada 1 hora a 37ºC con distintas concentraciones del mutante de tiorredoxina P34H y a 100µM de GSSG; en el caso de los puntos verdes con 14.4µM, los triángulos de borde azul con 9.66 µM y los rombos de borde negro con 2.35µM.

Posteriormente las cinéticas de hidrólisis del 2′, 3′ –cCPM fueron ajustadas a una ecuación cuadrática:

$$A_t = A_i + kt + ct^2 \qquad (25)$$

Donde A_t es la absorbancia a tiempo t, A_i es la absorbancia a tiempo = 0, y k será la velocidad inicial, es decir dA_t/dt cuando t=0. Las velocidades iniciales son representadas frente a las concentraciones de las enzimas con las que previamente fue incubada la ribonucleasa y que catalizaron su recuperación mediante la formación o isomerización de sus puentes disulfuro, Figura 23, y la pendiente del ajuste lineal de estos datos será la forma en la podamos

expresar dichas actividades recuperadoras (ya sea de formación o isomerización de puentes disulfuro).

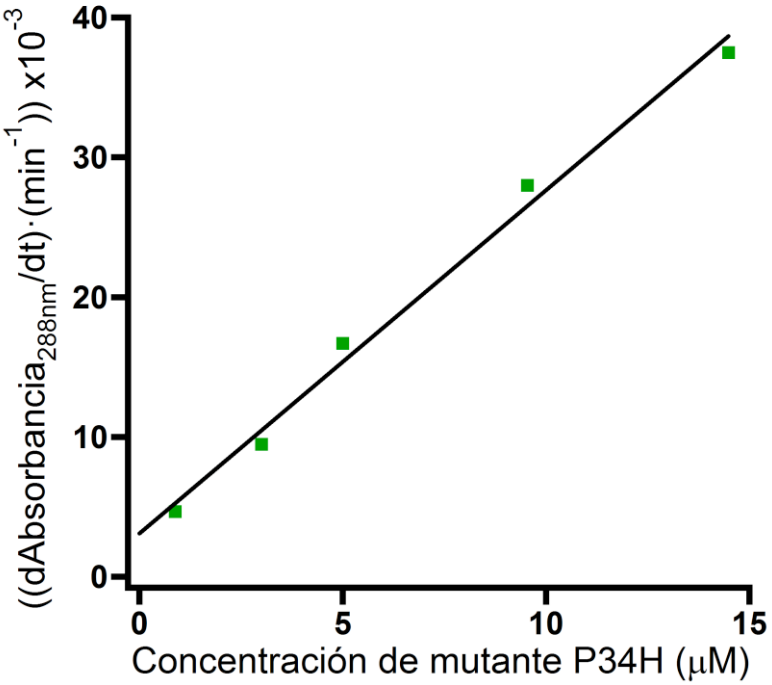

Figura 23. Valores de las velocidades iniciales de la hidrólisis del 2', 3' –cCMP representados frente a las concentraciones del mutante P34H que fueron incubados con RNasa reducida durante 1 hora a 37ºC a 100μM de Glutatión oxidado (GSSG). El valor de la pendiente es 0.002456 ± 0.000137 ($min^{-1}\cdot\mu M^{-1}$), será tomado para expresar la actividad formadora de puentes disulfuro del mutante P34H. El valor de la ordenada en el origen (0.003095 ± 0.0012 min^{-1}) es el de la formación de puentes disulfuro no catalítica por el GSSG.

2.4.3.2.1 Metodología experimental

2.4.3.2.2 Preparación de la ribonucleasa

2.4.3.2.2.1 Preparación de la RNAsa reducida

Fue realizado como está descrito en el artículo publicado por el grupo de Arne Holmgren [64]:

a. Se prepara una disolución de Ribonucleasa a 30 mg/mL en 9 M urea, 130 mM DTT, pH: 8.6 equilibrado con Tris-base y todo ello saturado de N_2 para tener unas condiciones lo menos oxidantes posibles.

b. Se incuba dicha disolución durante 60 min a 37ºC.

c. Se baja el pH de la disolución a 4 mediante la adición de ácido acético glacial.

d. Se cambia el tampón de la disolución:

I. Primero se pasa la muestra por una columna Sephadex G-25 (Pharmacia PD10), previamente equilibrada con ácido acético 0.1 M.

II. Después dializamos frente a 0.1 M de ácido acético 0.1 M durante 2 horas a 4ºC.

e. Por último se hacen alícuotas a 4 mg/mL y se guardan a -20ºC.

2.4.3.2.2.2 Preparación de la RNAsa con los puentes disulfuro mal formados. (RNasa scrambled)

La preparación de la ribonucleasa con los puentes disulfuro mal formados (scrambled), que es el sustrato para los ensayos de actividad isomerasa fueron preparados esencialmente como está descrito por Hillson *et al.* [68].

a. Tras pasar la ribonucleasa reducida por la columna Sephadex G-25 como se describe en el apartado anterior, esta se diluye hasta 0.5 mg/mL en 9 M urea, 0.1 M sarcosina-HCl.

b. Se ajusta el pH de la disolución a 8.5 adicionando Tris-base a 1 M.

c. Se deja oxidar la disolución a temperatura ambiente, a oscuras, en agitación y expuesta al aire, hasta que esta contenga menos de 0.1 mol de grupos tioles por mol de moléculas de RNAasa (aproximadamente a los 3 días).

d. El contenido de grupos tioles se determinó usando 1 mM ácido nitrobenzoico en 6 M de hidrocloruro de guanidinio, pH: 8.0 [69].

e. La muestra se dializa frente a 50 mM $(NH_4)_2CO_3$ y se liofiliza.

f. El liofilizado se disuelve en 0.1 M ácido acético hasta 4 mg/mL, se hacen alícuotas y se congela.

2.4.3.2.3 Procedimiento de los ensayos de actividad de formación e isomerización de puentes disulfuro

Los ensayos de actividad constan de dos partes, una primara fase de incubación, en la que la ribonucleasa sustrato es incubada con la proteína de la que se pretende testar su capacidad replegarla, bien catalizando la formación de puentes disulfuro, bien catalizando la isomerización de los mismos, y la segunda fase dirigida a conocer la cantidad de ribonucleasa que ha sido ''recuperada'' mediante el ensayo de la hidrólisis del 2', 3' –cCMP.

2.4.3.2.3.1 Incubación.

a. En un tubo Eppendorf de 1.5 mL se añaden en estas cantidades y orden:

I. 70 µL de la proteína a testar para alguna de las dos actividades, el tampón en el que se encontraban era HEPES 5mM, pH:7.

II. 380 µL de tampón fosfáto potásico 0.1M, 1mM EDTA, pH: 7 y 131.6 µM (para que la disolución final quede a 100µM) de GSSG en el caso del ensayo de formación de puentes disulfuro, o de GSH para el ensayo de isomerización. Este tampón ha de estar saturado de N_2.

III. 50 µL de la RNasa reducida o scrambled preparada anteriormente en 0.1M Ac. acético, para que la concentración final quede a 0.4 mg/mL.

b. Se incuba la mezcla a 37ºC durante una hora.

2.4.3.2.3.2 Ensayo hidrólisis 2', 3' –cCMP.

Se registra la hidrólisis espectrofotométricamente siguiendo la variación de la absorbancia a 288 nm con el tiempo a 25ºC.

a. Transcurrido el periodo de incubación descrito en el apartado anterior, se mezclan dentro de una cubeta de cuarzo:

I. 200 µL de la solución de incubación.

II. 800 µL de tampón 50 mM Tris-HCl, 25 mM KCl, 5 mM $MgCl_2$ a una concentración final de 200 µg/mL de 2', 3' –cCMP, a pH: 7.5

b. Rápidamente se mezcla el volumen de la cubeta con una pipeta pasteur y se inicia la media a 288 nm durante 15 minutos.

Dado que lo que nos interesa es obtener la velocidad inicial de esta cinética, se debe medir el tiempo que tarda en iniciar la medida el espectrofotómetro desde que iniciamos la reacción al añadir los 200µL de la solución de incubación.

Al igual que el ensayo de actividad reductasa se debe tener la precaución de no salir nunca del rango de concentraciones de catalizadores que guarden una relación lineal con las velocidades iniciales de hidrólisis, como ocurre en la Figura 23. Al realizar los ensayos de actividad de los mutantes de la biblioteca siempre se hicieron primero una serie de experimentos con el mutante de partida de dicha biblioteca, el mutante P34H, para comprobar que los datos experimentales obtenidos caen dentro de la recta patrón.

2.4.4 Electroforesis en geles de acrilamida (PAGE) nativa.

La electroforesis nativa es una técnica electroforética que sirve para separar proteínas en condiciones nativas, se realiza por tanto en ausencia de agentes desnaturalizantes como el SDS (siglas en inglés de dodecil sulfato sódico). Las proteínas serán separadas por su tamaño y carga neta de su forma nativa a un determinado pH.

En este trabajo esta técnica fue utilizada exclusivamente para determinar la diferencia de carga neta de las proteínas del Artículo 1 respecto de la variante trx* (D10A/A22P/ I23V/Q50A/P68A/G74S/E85Q/A87V) [34] a pH neutro, pH al que se realizó toda su caracterización. Como todas las proteínas que fueron comparadas fueron variantes de tiorredoxina con cola de histidina, y tenían consecuentemente aproximadamente el mismo tamaño, las diferencias vistas en el gel entre las distintas variantes tan solo van a corresponder a diferencias de carga neta entre ellas.

2.4.4.1 Metodología experimental

La electroforesis nativa fue realizada de acuerdo con el protocolo comercial (Instruction Manual Mini-PROTEAN® Tetra Cell de BIO-RAD). Se utilizó un sistema de tampón continuo, es decir que el tampón usado para preparar el gel es el mismo que el tampón de desarrollo y también es usado para preparar la muestra, el gel por tanto carece de la parte de apilamiento. El tampón del sistema continuo fue 43 mM Imidazol, 35 mM HEPES (forma ácida) pH: 7.4 de acuerdo con Mc Lelan [70].

Los geles utilizados se prepararon al 6 % de poliacrilamida y la concentración final de proteína cargada en los geles fue de aproximadamente 0.6 mg/mL.

Para poder asignar una carga neta relativa a las variantes de la biblioteca del artículo incluido en el apartado 3.1 respecto a la trx*, ya que en este artículo se trabaja con proteínas que tienen residuos ionizables que por su condición de enterrados no siempre pueden encontrarse cargados, se utilizaron una serie de variantes de trx* que contenían mutaciones que añadían cargas a la superficie de la proteína a modo de marcador, estas fueron:

T14K/T54K/A105K (+3 respecto trx*)

H6E/T14K/T54K/Q62K/Q85K/Q98E/A105K (+3 respecto trx*)

T14K/T54K/N83D/Q85K/Q98E (+1 respecto trx*)

A68E/N83D/A105K (-1 respecto trx*)

A50E/A68E/N83D/Q85K (-2 respecto trx*)

Por último cabe citar que los geles fueron teñidos con tinción de Coomasie, siguiendo el protocolo convencional.

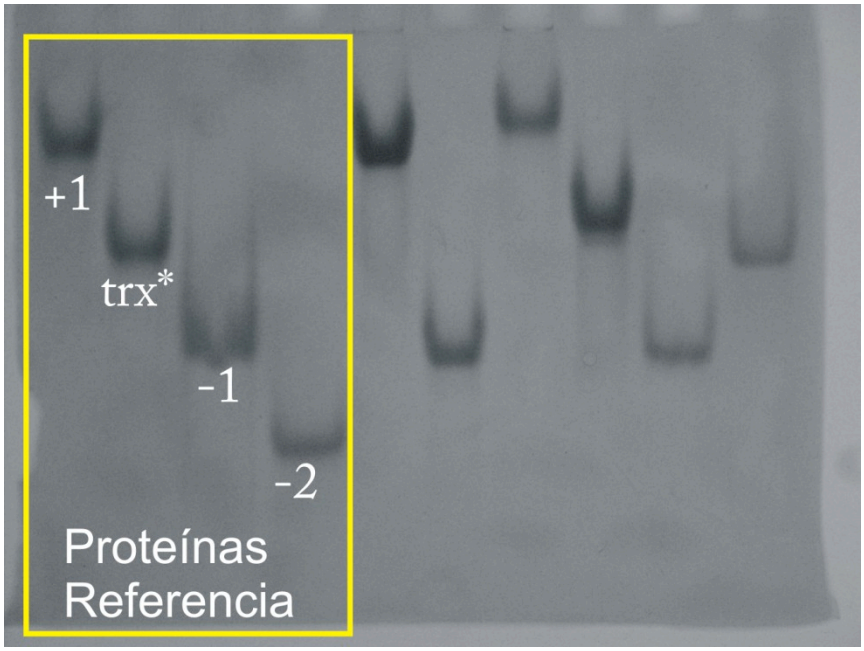

Figura 24. Gel PAGE nativo para determinar la carga neta de las 6 variantes de la derecha respecto a la variante de partida de la biblioteca trx*. Las variantes de la derecha de gel son las variantes (de izq. a derecha): 1, 2, 6, 7, 12 y 13 de artículo descrito en el apartado 3.1. Los números de la izquierda indican la carga neta de las proteínas de referencia respecto a trx*, dichas proteínas son: (+1) que corresponde al mutante T14K/T54K/N83D/Q85K/Q98E sobre trx*, (-1) es A68E/N83D/ A105K sobre trx* y (-2) es A50E/A68E/N83D/Q85K sobre trx*. Mientras que las 6 variantes de la derecha corresponden a los mutantes: (de izquierda a derecha) L7K, I72E, L58E/L79K, I72K/L79E, V16K/I72E/L79E y V16E/L58E/L79K todos sobre trx*.

2.5 PLS-R (Partial Least Squares Regression)

2.5.1 Introducción

La regresión por mínimos cuadrados parciales, PLS-R, es un método de análisis multivariante utilizado para determinar el efecto que tienen un grupo de predictores (variables independientes o variables **x**) sobre una o varias variables respuesta (variables dependientes o **y**). Llamaremos **Y** a la matriz

formada por el conjunto de M variables **y** de N muestras, y **X** a la matriz de K variables **x** de N muestras (ver Figura 25). Siendo el objeto del PLS-R, como método de regresión que es, el de predecir **Y** de **X**, este es particularmente adecuado en el caso de que el número de variables **x** sea muy grande en comparación con las variables **y**, e incluso que superen en número a las muestras (K > N). También es resaltable el hecho de que sea capaz de trabajar con datos con considerable ruido, colinearidad e incluso con una elevada proporción de datos incompletos.

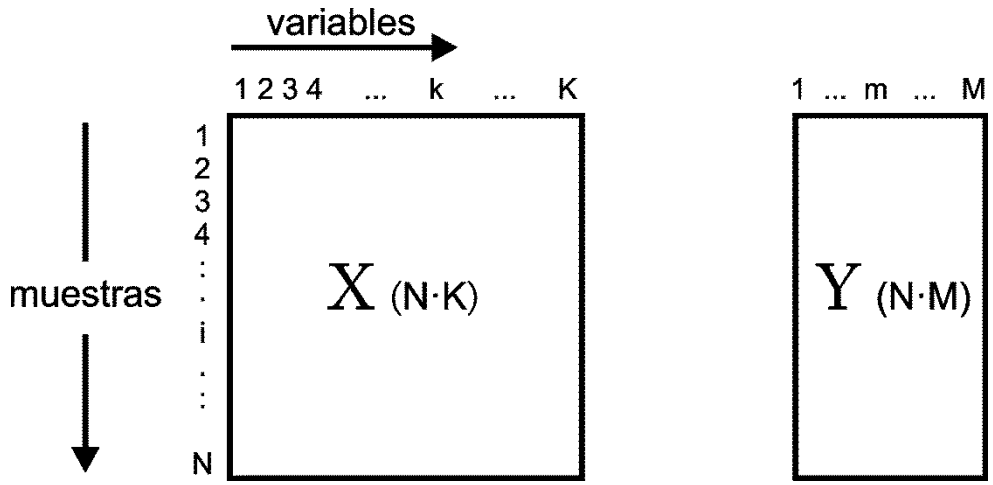

Figura 25. El PLS-R trabajará con un conjunto de N observaciones, o muestras estudiadas con K variables independientes x, y M variables independientes y. Formando así dos matrices X de dimensiones (N·K) e Y (N·M).

La investigación en ciencia, ingeniería así como en las ciencias sociales, normalmente implica inferir o explicar el comportamiento algunas variables dependientes (respuesta) a partir de otra o varias variables independientes (predictoras o explicativas) que sean fáciles de medir. A veces una sola variable respuesta puede ser explicada por una sola variable explicativa, dando lugar a una relación lineal que puede ser explicada por la ecuación de la recta del tipo:

$$y=b_0+bx+e \qquad (26)$$

Donde y es la variable dependiente, x es la variable independiente, b_0 el término independiente, b el coeficiente de regresión y e el error residual. Pero

no todas las relaciones son tan simples, y frecuentemente ocurre que una variable dependiente no puede ser explicada por una sola variable independiente, si no que necesita varias de ellas. La forma más sencilla de abordar este problema es el uso de la Regresión Lineal Múltiple (MLR, por sus siglas en inglés), técnica multivariante que relaciona de forma lineal una o más variables dependientes con las variables independientes mediante la siguiente expresión:

$$y=b_0+\sum_{k=1}^{K} b_k x_k+e \tag{27}$$

Donde x_k son las variables independientes, b_k son los coeficientes de la regresión y e el error residual. Sin embargo la MLR presenta una serie de limitaciones, como que el número de muestras nunca puede superar al número de variables x, puesto que encontraríamos un infinito número de soluciones de b [71]; además las variables independientes deben ser pocas y linealmente independientes entre ellas, es decir, que no se puedan expresar como combinaciones lineales de otras variables independientes, de lo contrario este método multivariante se vuelve bastante ineficiente.

Para soslayar el problema de la colinearidad existen diversos tipos de aproximaciones, los dos métodos más populares son los esquematizados en Figura 26. Uno sería utilizar la MLR con solo un número pequeño de variables independientes previamente seleccionadas, obviando el resto de variables (con la pérdida de información relevante que puede conllevar). Y un segundo tipo de métodos, llamados métodos basados en la reducción de variables, en el que las variables originales se combinan para dar lugar a unas pocas variables linealmente independientes (ortogonales), también llamadas factores.

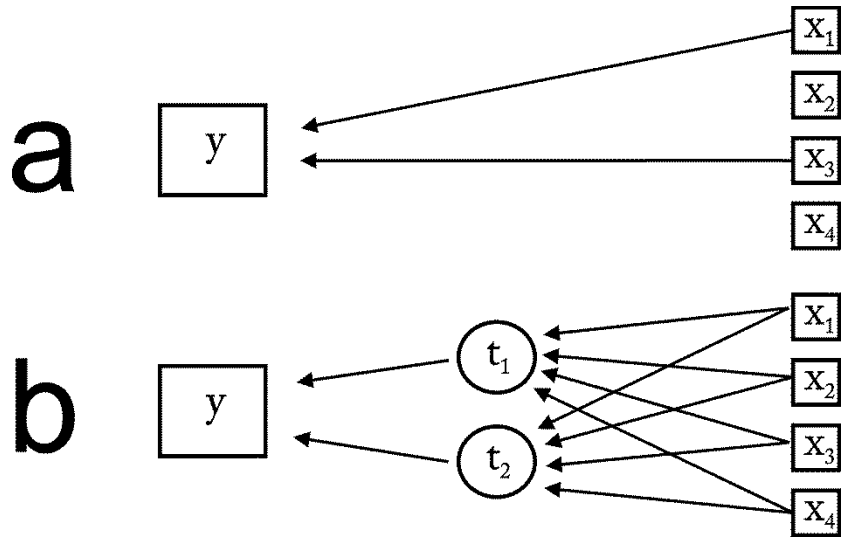

Figura 26. Diagrama conceptual de los dos tipos de métodos más populares en el tratamiento de la colinearidad. En **a** se seleccionan cuidadosamente unas pocas variables, prescindiendo del resto, para construir el modelo mediante MLR. En **b** las variables x se combinan linealmente para dar otro tipo de variables (t_1, t_2) que son menos en número que las originales y linealmente independientes entre sí, de manera que estas nuevas variables inferidas se relacionarán con y mediante una ecuación de regresión [72].

De entre los métodos basados en la reducción de variables cabe destacar dos, PCR (Regresión de Componentes Principales) y el PLS-R.

El PCR es un método de regresión múltiple basado en el análisis de componentes principales PCA, por sus siglas en inglés. El PCA es una técnica de análisis multivariante utilizada para reducir la dimensionalidad de un conjunto de datos. Partiendo de una matriz X de dimensiones (N·K), donde N es el número de observaciones y K el número de variables x, el PCA encontrará un número menor de variables ortogonales t, que resulten de combinar linealmente las variables originales x. Estas variables t, también se conocen como componentes o factores principales. En realidad el PCA selecciona un nuevo sistema de coordenadas para el elenco de datos, de forma que el primer eje, primer componente principal sea la dirección que explique la mayor varianza de estos datos, el segundo componente explicará

la mayor varianza que no ha sido explicada por el primer componente, siendo de esta forma ortogonal a él, y así sucesivamente.

Por tanto el conjunto de variables t resultante contendrá la información más relevante de la matriz X eliminando redundancias y ruido. Matemáticamente podemos expresar entonces X como:

$$X=t_1p_1^{'}+t_2p_2^{'}+...+t_ap_a^{'}+e \quad (28)$$

Siendo los vectores t_a (llamados también ''scores'') cada componente principal, los vectores p_a los ''loadings'' que describen la relación entre t y las variables originales x, y siendo el vector e residual que representa el ruido y la varianza irrelevante de X. O en términos de matrices:

$$X=TP'+E \quad (29)$$

T es la matriz de componentes principales, P' la matriz de loadings y E la matriz de residuales de X.

Para tener una reproducción exacta de la matriz X, *a* debe ser igual a K, pero dado que la información más relevante, mayores varianzas, está contenida en los primeros componentes, podremos ''resumir'' la información de X con un número de componentes *a* bastante menor que K. La regresión por componentes principales PCR, utiliza las nuevas variables t encontradas por el PCA para predecir las variables y, de la misma forma que MLR lo hacía con las variables originales; de esta forma podemos reescribir su la ecuación de regresión así:

$$Y=TB+F \quad (30)$$

Donde B será la matriz de los coeficientes de regresión, y F la matriz de los residuales de Y que expresa la desviación de los valores reales sobre los predichos por el modelo.

El problema es que las nuevas variables seleccionadas para la regresión son las que mejor representan la matriz X, pero en principio no hay motivos para pensar que vayan a ser estos buenos predictores de Y, ya que puede existir parte de variabilidad en X que no correlacione con Y; o que gran parte de la

variabilidad que correlaciona con Y se esté desechando por encontrarse entre los últimos componentes, que son los que menos explicarían la variabilidad en X. Por este motivo se han desarrollado otras técnicas de regresión como el PLS-R.

La regresión parcial por mínimos cuadrados, PLS-R, fue desarrollada por el economista Herman Wold en 1975 [73]. Originariamente fue concebida para su uso en el campo de la econometría, aunque posteriormente se convirtiera en una técnica muy popular en la quimiometría [71, 74], y hoy en día se haya extendido a otras muchas áreas del conocimiento. El PLS-R a diferencia de PCR utiliza ambas matrices la de X e Y (ya que como en las otras técnicas de regresión se pueden modelar varias y a la vez) en la estimación de los componentes, de manera que se intenta que los primeros componentes sean los que más influencia predictiva tengan sobre la matriz Y.

2.5.2 Modelo PLS-R

En el modelo de PLS-R las matrices de X e Y también pueden descomponerse, al modo del PCA, en una serie de ''scores'' y de ''loadings'' ver Figura 27, donde los scores serían nuevas variables inferidas por el modelo y los loadings los coeficientes regresores con que se relacionan unas variables con otras.

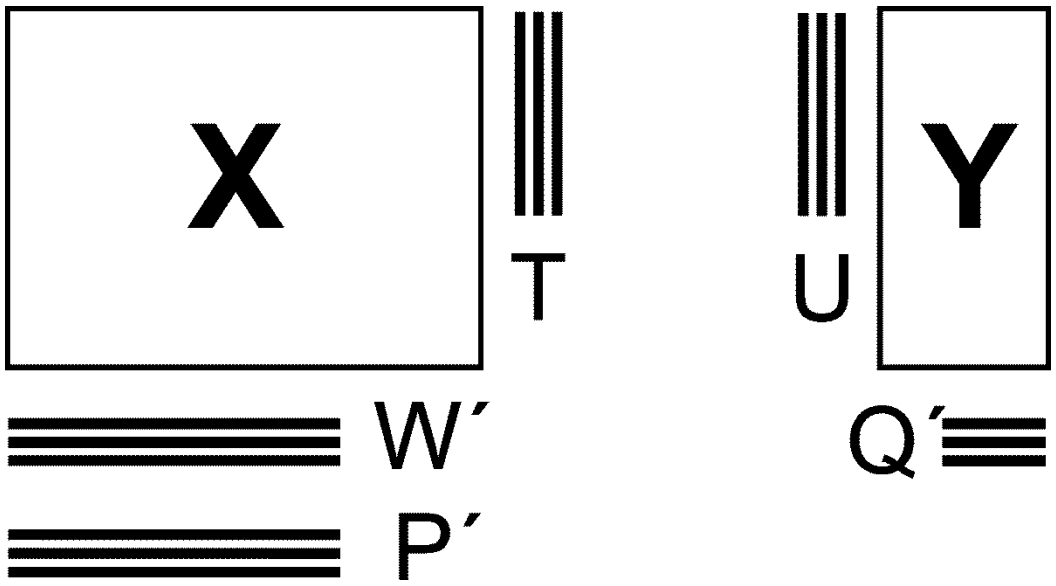

Figura 27. Esquema de PLS-R. T son las nuevas variables o factores, se diferencian de los componentes del PCA en que estos están elegidos para captar la parte de X que mejor prediga Y. U que no existen en PCA resumen la parte de Y que es explicada por X para un factor dado. Siguiendo con la terminología del PCA hay dos tipos de loadings de x, P que determina cuanto contribuye cada variable x a cada variable t, W para expresar cuanto contribuye cada variable x a explicar Y, y un loading de y Q para explicar la relación entre las variables y y las t. [75]

En el modelo PLS-R las nuevas variables t también son pocas y ortogonales entre sí como en el PCA, pero estas son estimadas como combinaciones lineales de las variables x con unos nuevos coeficientes W llamados ''weights'' (pesos); estos coeficientes que no están presentes en el análisis PCA, expresan lo que contribuye cada variable x a explicar Y de forma que podamos expresar, en términos de matrices:

$$T=XW' \tag{31}$$

Podemos descomponer las matrices X e Y en:

$$X=TP'+E \tag{32}$$

$$Y=UQ'+G \tag{33}$$

Donde E y G serían las matrices de residuales de X e Y respectivamente, que deben ser pequeñas. Pero esta descomposición de las matrices X e Y no es

independiente, sino que es concurrente y está interrelacionada, de forma que se cumpla la relación para cada componente a:

$$u_a = h_a t_a \quad (34)$$

Donde u y t, están relacionadas por una constante h para cada a componentes. La relación entre u y t se puede considerar una síntesis de la relación entre las matrices X e Y para cada componente. Ya que las variables t son escogidas para ser buenas predictoras de Y:

$$Y = TQ + F \quad (35)$$

La matriz F de residuales de Y expresa la desviación de los valores reales sobre los predichos por el modelo, se pretende que ésta sea lo menor posible.

Sustituyendo la ecuación (31) en (35), podemos llegar a una ecuación de regresión múltiple del siguiente tipo:

$$Y = XWQ + F = XB + F \quad (36)$$

Donde B será la matriz de coeficientes de regresión, que es escrita como B=WQ.

Tras el cálculo del primer componente las matrices X e Y se ''desinflan'', es decir se les sustrae respectivamente t_1p_1 y t_1q_1, lo que equivale a quitarles la información explicada por el primer componente, de forma que el segundo componente se calculará sobre los residuales del primer componente e_1 y f_1, así sucesivamente para el resto de componentes hasta que se haya extraído el número deseado de estos.

Si el número de componentes a es igual o mayor al rango de la matriz, entonces PLS equivaldría a MLR.

El algoritmo utilizado en este trabajo para calcular el modelo PLS-R fue NIPALS (Nonlinear Iterative Partial Least Squares) que es comúnmente utilizado también en la técnica de PCA. Este algoritmo es capaz de tolerar cierta cantidad de datos incompletos, lo hace sustituyendo iterativamente los datos vacios por predicciones.

2.5.3 Validación del modelo

Todo modelo, ya esté destinado a la explicación de datos experimentales o a predecir nuevos, requiere previamente ser validado. En el caso que nos concierne, el de la regresión PLS, la validación consistirá en la evaluación de la capacidad predictiva del modelo, de manera que si esta es razonablemente buena el modelo se dará por válido. La calibración del modelo obtendrá unos parámetros de regresión que nos permitirán predecir el comportamiento de una o varias variables respuesta, de forma que para cada muestra i y variable dependiente m se pretenderá que su residual f_{im} sea lo más pequeño posible:

$$f_{im}=\hat{y}_{im}-y_{im} \quad (37)$$

Donde $\hat{y}_{im}$ es el valor calculado por el modelo y y_{im} es el valor ''real'' (verificado experimentalmente).

Hay varios métodos que pueden usarse para estimar la bondad de nuestro modelo. Lo ideal es disponer de un conjunto de datos independientes al modelo, que sean conocidos, pero que no hayan sido utilizados para su cálculo, de manera que podamos evaluar la capacidad del modelo para predecirlos. El problema es que, como es nuestro caso, no siempre se dispone de una cantidad suficiente de datos experimentales para calibrar el modelo con un subgrupo de ellos y validarlo con otro. Por tanto, una buena alternativa es usar el método de validación cruzada, que es un método llamado de validación interna debido a que los datos que utiliza para validar el modelo son los mismos que usan para calibrarlo. En la validación cruzada el conjunto de datos se divide en varios subconjuntos o segmentos, de manera que secuencialmente se va prescindiendo de un segmento cada vez para realizar la calibración, construyendo el modelo con el resto, mientras que el segmento que se deja fuera será predicho por el modelo y por tanto servirá para validarlo. Se trata de que finalmente todos los segmentos contribuyan a la calibración y a la validación del modelo. La validación de un modelo se hace para cada grado de complejidad de este, es decir, se realizará la

validación para el modelo explicado por un solo componente principal, luego se validará tras incluir el segundo componente y así sucesivamente hasta incluir todos los componentes. De manera que tendremos el modelo validado para cada número de componentes que lo conformen. Existen varios tipos de validación cruzada, nosotros utilizamos la validación cruzada completa, donde el número de segmentos es igual al número de muestras N, por lo que cada vez se deja una sola muestra fuera. El error de validación, que será la varianza residual de Y, para cada número de componentes viene dado por la siguiente expresión:

$$\mathrm{MSEP}=\frac{\sum_{i=1}^{j}(\hat{y}_{im}-y_{im})^2}{j} \tag{38}$$

Esta expresión es la media de la suma de cuadrados del error de predicción, MSEP por sus siglas en inglés, donde $\hat{y}$ son los valores predichos e y los experimentales, j será el número de segmentos utilizados en la validación, que en nuestro caso será igual a N muestras.

2.5.4 Número óptimo de componentes

Una de las cuestiones clave en este tipo de modelos de reducción de variables, es cómo seleccionar el número óptimo de componentes que formarán parte de nuestro modelo. Los datos experimentales siempre llevan asociado cierto ruido, por pequeño que sea, y por tanto es de esperar que algunos de los últimos componentes solamente estén explicando la contribución de este ruido. También se suele eludir el uso de estos últimos componentes para evitar el problema de la colinearidad, por tanto la inclusión de demasiados componentes supondrá un sobre-ajuste del modelo, que le hará perder capacidad predictiva. Pero, si por el contrario, usamos demasiados pocos componentes obtendremos un modelo que no será capaz de explicar parte importante de la variabilidad de los datos, lo que llamamos sub-ajuste. Por tanto se tratará de encontrar este óptimo número de

componentes que nos dé la mayor capacidad predictiva del modelo. En la Figura 28 se ilustran ambos problemas.

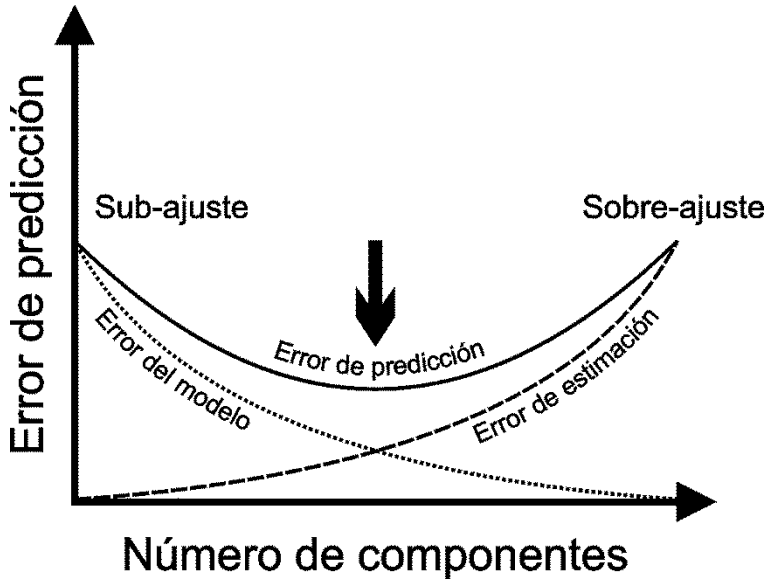

Figura 28. Gráfico de error de predicción frente al número de componentes que conforman un modelo de reducción de variables como el PLS [76]. A medida que el número de componentes usados por el modelo aumenta, también aumenta el error de estimación, ya que mayor número de parámetros han de ser estimados; viéndose por ende afectado el poder predictivo del modelo. Por el contrario a mayor número de factores menor será el error del modelo ya que mayor parte de la variabilidad de los datos estará calibrada.

Para tratar de encontrar la correcta complejidad del modelo dotándole del número óptimo de componentes, el programa The Unscrambler 10.1 (CAMO Software), que fue el utilizado para realizar la regresión PLS, propone elegir el número de componentes en base a la siguiente regla [75]:

$$\mathrm{Min}[\mathrm{Vytot\_val}_{PC=0}\cdot 0.01\cdot a+\mathrm{Vytot\_val}_{PC=a}] \quad (39)$$

Se elegirá el número de componentes a los que corresponda el mínimo del cálculo de la expresión entre corchetes, donde a es un número dado de componentes, Vytot_val es la varianza residual de validación de y, que para un número de componentes de cero, PC=0, equivale a la varianza inicial de *y*, mientras que para un número de componentes a, PC=a, corresponderá a MSEP para a número de factores.

Es decir para cada nuevo componente principal se le añade un 1% de la varianza inicial a la varianza residual correspondiente a un número de componentes a, evitando de esta forma elegir un número de componentes que no mejore sustancialmente la bondad de predicción y así evitando incurrir en un sobreajuste del modelo.

2.5.5 Procedimiento

En este trabajo usamos la herramienta de regresión PLS-R para el estudio de bibliotecas combinatoriales de mutantes, un procedimiento novedoso [77-79] y con un gran potencial en el campo de la ingeniería de proteínas. En este trabajo se trató de explicar el comportamiento de una serie de propiedades de las proteínas de una biblioteca combinatorial, tales como actividades enzimáticas y estabilidad térmica, que serían las variables dependientes, relacionándolas con un gran número de variables independientes que contenían información de la secuencias de las proteínas, así como de las posibles interacciones entre residuos. Esta relación se representa en la siguiente ecuación:

$$\ln y^{m} = \sum_{k} x_{k} \cdot p_{k}^{m} + \sum_{k} \sum_{l \neq k} x_{kl} \cdot p_{kl}^{m} \qquad (40)$$

Donde y^m sería una variable dependiente m, x_k es una variable independiente que tomará valores de 0 o 1 en función de la ausencia o presencia, correspondientemente, de mutación en la posición k, p_k^m será un coeficiente de regresión que expresará el efecto de una mutación en la posición k sobre la actividad y^m, $x_{kl} = x_k \cdot x_l$ será una variable independiente que tomará valor de 1 si las mutaciones en k y l ocurren simultáneamente y en caso contrario será 0, p_{kl}^m será el coeficiente de regresión que mida el efecto acoplado de las mutaciones de las posiciones k y l sobre y^m. En esta ecuación se representa un modelo complejo en el que se ven involucradas un elevado número de variables independientes, que de hecho son superiores en número

a los datos que se pretenden ajustar, problema que abordamos mediante el uso de la herramienta PLS-R que, como se ha explicado en este apartado, será capaz de realizar la regresión usando un número muy reducido de nuevas variables o componentes.

2.5.5.1 Tratamiento previo de los datos

Antes de desarrollar el modelo de PLS-R es muy recomendable realizar un tratamiento previo de los datos para asegurar que todas las variables de un bloque, ya sea X o Y, tengan el mismo peso a la hora de hacer el análisis, el tratamiento seguido en este trabajo fue el siguiente:

a. Las variables *y* que tenían una variabilidad de datos superior al orden de magnitud las pasamos a logaritmo neperiano.

b. Realizamos posteriormente lo que se denomina un auto-escalado de los datos, en el que se primero se centra cada variable sustrayendo su media y posteriormente se divide por su desviación estándar.

Desde el punto de vista geométrico, este tratamiento equivaldría a igualar la longitud de los ejes de coordenadas para que puedan equipararse así los pesos de las distintas variables, que tenían distintas unidades, en el cálculo del modelo.

En nuestro caso solamente realizamos este tratamiento sobre las variables *y*, ya que al ser datos bastante heterogéneos, (provienen de la caracterización de diferentes actividades y estabilidad de los mutantes), serían sino incomparables unas con otras, ya que tendrían escalas y unidades diferentes.

2.5.5.2 Bootstrapping

Dado que para la construcción del modelo solo se disponían de los datos del estudio de 30-40 mutantes de los 1024 que contiene la biblioteca combinatorial, este estará expuesto a una gran dosis de incertidumbre ya que puede verse seriamente condicionado por la presencia de datos anómalos

debidos a errores experimentales, también conocidos como ''outliers'', que no siempre son fáciles de detectar. Por este motivo decidimos hacer un remuestreo de nuestro conjunto de datos usando la técnica llamada bootstrapping, propuesta por Bradley Efron en 1979 [80]. Esta técnica se basa en la idea de que cuando no se dispone de más información sobre una población que la de un pequeño conjunto de datos aleatorios, ésta es la mejor pista de que disponemos para conocer la distribución de la población total, de forma que remuestrear sobre dicho conjunto aleatorio será lo más cercano a remuestrear sobre toda la población [81]. El remuestreo de los datos se realiza de forma aleatoria con reemplazamiento, por tanto si tenemos un conjunto de N muestras estudiadas, cada una de las muestras de bootstrapping, o réplicas, tendrá también tamaño N.

Para cada conjunto de datos experimentales obtuvimos 20 réplicas, con cada una de ellas se construyó un modelo de PLS-R.

2.5.5.3 Predicción y elección del set de Pareto

La regresión PLS tiene como fin último la predicción de nuevos datos a partir de los datos ya conocidos. Así una vez que construimos el modelo con un conjunto de datos experimentales, podemos predecir el comportamiento de toda la población, es decir, en nuestro caso de todos los mutantes de la biblioteca combinatorial. Se hizo una reconstrucción de toda la biblioteca a partir de cada una de las 20 réplicas, en la Figura 29 se ilustra a modo de explicación la reconstrucción de 3 réplicas.

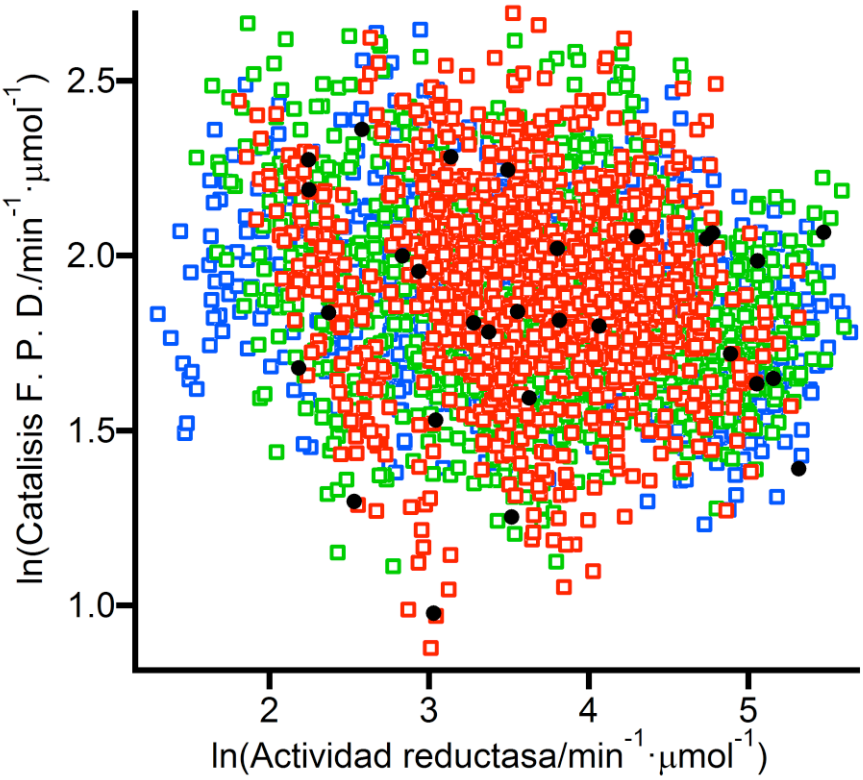

Figura 29. Reconstrucción de los valores actividad de catálisis de formación de puentes disulfuro y de actividad reductasa para toda la biblioteca combinatorial, de 1024 variantes. En círculos negros se hayan representados los datos experimentales, mientras que los cuadrados corresponden a las reconstrucciones de la biblioteca entera a partir de tres réplicas distintas del conjunto de datos experimentales, cuadrados de bordes verde, azul y rojo. Para la reconstrucción se utilizaron los modelos de PLS-R con el número de componentes óptimo, calculados según se describe en el apartado anterior.

Cuando consideramos la optimización simultánea de varias propiedades, en concreto de dos como es el caso del presente trabajo, no es posible definir una única solución, si no un conjunto de soluciones optimas, que pueden considerarse como el conjunto de soluciones no dominadas o frontera de Pareto, entendiendo que una variante **a** domina sobre una variante **b** si y solo si **a** mejora o iguala simultáneamente los valores de dos propiedades de **b** [36], así las soluciones presentes en la frontera de Pareto no son dominadas por ninguna otra solución. Esta estrategia muy usada en economía, está recientemente siendo usada en el campo de la ingeniería de proteínas [9, 36,

37] y de la biología evolutiva [35]. En la Figura 30 se ilustra la frontera de Pareto de las reconstrucciones de la Figura 29.

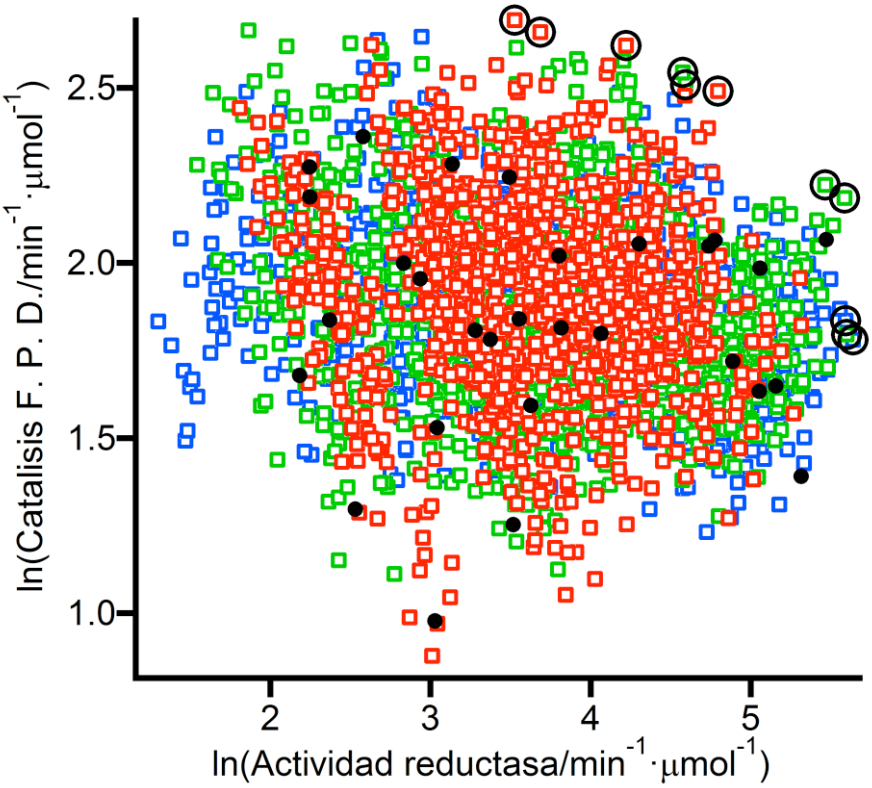

Figura 30. La frontera de Pareto de las reconstrucciones de la biblioteca de la Figura 29, está representado por los círculos huecos negros, que engloban las soluciones no dominadas del conjunto de datos formado por las tres reconstrucciones de la biblioteca combinatorial.

Para encontrar las potenciales mejores variantes de la biblioteca, soluciones óptimas, calculamos la frontera de Pareto del conjunto de las reconstrucciones de los modelos PLS-R pertenecientes a las 20 réplicas de cada conjunto de datos experimentales. Posteriormente estas soluciones elegidas se estudian experimentalmente, lo cual es la mejor forma de validar el modelo, y a su vez pasarán a formar parte del conjunto de datos experimentales sobre los que podremos repetir todo el proceso que acabamos de explicar, de este modo con más datos experimentales es de esperar obtener un mejor modelo que se

ajuste mas a la realidad y que por tanto sea capaz de realizar mejores predicciones.

2.6 Análisis estadístico de secuencias

El estudio de la secuencia aminoacídica de las proteínas, la estructura primaria, ha sido vital en el desarrollo de la ingeniería de proteínas en los últimos años, la introducción de estudios comparativos entre secuencias han sido de acusada importancia para la predicción de estructuras, funciones, contactos tridimensionales entre proteínas y un largo etcétera. Las secuencias albergan una valiosa información, sobre el historial evolutivo de las proteínas, una "memoria", así comparando secuencias podemos detectar proteínas homólogas, que son aquellas que tienen el mismo antecesor común, ya que estas conservan una elevada proporción de similitud en sus secuencias.

Un análisis estadístico de secuencias de proteínas homólogas puede darnos información sobre las frecuencias de aparición de los residuos en las distintas posiciones de proteínas de una misma familia, de manera que se puedan localizar las posiciones más conservadas, se puede comprobar también cómo de favorecidas, desde el punto de vista evolutivo, estarían algunas mutaciones en las que estemos interesados, realizar análisis de correlación, análisis filogenéticos (reconstrucción de proteínas ancestrales)...

En el presente trabajo se realizaron algunos de los análisis estadísticos citados en el párrafo anterior, sobre un alineamiento de secuencias realizado sobre la secuencia problema de tiorredoxina de *E. coli*.

2.6.1 Alineamientos de secuencias

Un alineamiento de secuencias es un conjunto de secuencias ordenadas en base a la identidad que comparten con la secuencia de la proteína de nuestro interés, a la que llamaremos secuencia problema, entendiendo por identidad

de una secuencia problema A con una B el número de coincidencias entre ambas dividido por el número de residuos de la secuencia A.

Los alineamientos de secuencias realizados en este trabajo se hicieron tomando como secuencia problema la de tiorredoxina de *E. coli*. Se utilizó el programa WU-BLAST2 (http://blast.wustl.edu) para realizar búsquedas sobre las bases de datos de secuencias de Uniprot (TrEMBL y Swiss-Prot (http://www.ebi.ac.uk)). El WU-BLAST2 realiza el alineamiento respecto a la secuencia problema utilizando el algoritmo de Smith-Waterman [82]. Las secuencias que conforman el alineamiento son filtradas por un límite de identidad, desechándose aquellas con una identidad menor al 25%-30%, está considerado que las proteínas que comparten una identidad de al menos el 25% tienen el mismo plegamiento [83], salvo contadas excepciones, este corte de identidad nos garantizará que la inmensa mayoría de las secuencias que queden en el alineamiento pertenezcan a proteínas homologas a la proteína problema, y de esta manera podamos realizar sobre este alineamiento de secuencias los análisis estadísticos que nos disponemos a exponer.

2.6.2 Hipótesis de pseudo-equilibrio.

Las proteínas actuales son el resultado de miles de millones de años de evolución, a lo largo de la misma, las secuencias de las proteínas han ido sufriendo mutaciones que debido a la presencia de determinados umbrales fenotípicos, que actúan como límites selectivos, han podido ser toleradas o rechazadas. Ha sido sugerido en diferentes trabajos [83-86] que bajo una escala de tiempo suficientemente amplia como es la evolutiva, se puede considerar que los distintos residuos pueblan una determinada posición en la secuencia de una proteína de acuerdo al efecto beneficioso o pernicioso (de acuerdo con los límites selectivos) que éstos confieran a la proteína. De manera que si asignáramos a este efecto un valor de energía libre (de valor proporcional al beneficio que produce) podríamos considerar que el sistema

se comporta de manera análoga a una distribución de Boltzmann de energías en un sistema termodinámico en equilibrio, de forma que la probabilidad de encontrar un residuo x en una determinada posición podrá calcularse como:

$$p(x)=e^{\left(\frac{-\Delta G}{RT}\right)} \tag{41}$$

Así cada residuo x para una posición tiene asociada una energía libre estadística, R es la constante universal de los gases ideales y T la temperatura en Kelvin, que en este caso se puede considerar una medida de la presión evolutiva [83]. De modo que a mayor frecuencia de aparición del residuo x en el alineamiento de secuencias mayor energía, se ha comprobado que esta energía estadística correlaciona con la energía libre del desplegamiento experimental [84, 85], pudiéndose por tanto, basándonos en la ecuación (41) relacionar el efecto de una mutación, dada la frecuencia de aparición de los residuos en el alineamiento, con la estabilidad de una proteína utilizando la siguiente expresión:

$$\frac{N_b}{N_a}=e^{\left(\frac{-\Delta\Delta G_{a\to b}}{RT}\right)} \tag{42}$$

O reorganizando los términos:

$$\Delta\Delta G_{a\to b}=RT\cdot\ln\frac{N_a}{N_b} \tag{43}$$

Donde N es el número de residuos tipo a o b presentes en el alineamiento de secuencias para una determinada posición del alineamiento de secuencias.
Para los cálculos realizados en este trabajo se consideró T=300 K.

2.6.3 Análisis de correlación o acoplamiento

Este tipo de análisis detecta en qué medida dos posiciones de la secuencia de una proteína se encuentran evolutivamente acopladas, es decir, que el cambio (o perturbación) de un residuo en una posición A pueda implicar el cambio en otra posición B, este acoplamiento es evidenciado en los alineamientos de

secuencias como un cambio en las frecuencias de aparición de residuos en la posición B en respuesta al cambio de residuo en la posición A.

El estudio del acoplamiento evolutivo entre posiciones de una secuencia reviste un enorme interés, pone de manifiesto que las posiciones acopladas han co-evolucionado, este tipo de análisis ha permitido localizar redes de residuos interconectados en proteínas con involucrados en el plegamiento [87] y con importantes roles funcionales como la transmisión de señales, regulación alostérica y catálisis [88-93].

En este trabajo se realizaron los siguientes análisis de correlación:

2.6.3.1 Análisis estadístico de acoplamiento (SCA)

Este análisis, también conocido como análisis de perturbación, fue desarrollado por Lockless y Ranganathan [94]. Está basado dos conceptos:

- El de conservación, la conservación de una posición en un alineamiento de secuencias viene dada por la desviación de la frecuencia de los residuos en esa posición respecto de su frecuencia de aparición media en todas las proteínas. Por tanto si esta posición no estuviera condicionada evolutivamente la distribución de frecuencias de sus residuos no debería desviarse significativamente de la frecuencia de aparición media de estos.
- El acoplamiento estadístico entre dos posiciones j e i se define como el nivel en el que cambian las frecuencias de una posición j frente a un cambio o perturbación en i.

En este análisis se compararan la conservación, en forma de energía estadística, de una posición j del alineamiento total, con su conservación en un sub-alineamiento que contenga el cambio en i. De manera que la existencia de acoplamiento entre dos posiciones implicaría un mutuo condicionamiento evolutivo de las dos posiciones, una co-evolución.

Los detalles del análisis se describen a continuación:

En primer lugar se analiza la frecuencia de aparición de cada aminoácido para cada una de la posiciones del alineamiento de secuencias, de forma que una posición *j* dentro del alineamiento pueda ser descrita como un vector $\vec{f}_j$ compuesto por la distribución de frecuencias que presenta cada aminoácido en esa posición *j*.

$$\vec{f}_j=\left\{f_j^{\,ala}, f_j^{\,cys}, f_j^{\,asp},..., f_j^{\,tyr}\right\} \tag{44}$$

Cada componente del vector $\vec{f}_j$ de frecuencias puede ser convertido a su probabilidad binomial, que es la probabilidad de encontrar un aminoácido *x* en una posición *j* dada la frecuencia de aparición media de este aminoácido en todas las proteínas de la base de datos. Para ello se utiliza la función de densidad binomial:

$$P(x)=\frac{N!}{n_x!(N-n_x)!}\, p_x^{n_x}\,(1-p_x)^{N-n_x} \tag{45}$$

En la que N será el número total de secuencias de alineamiento, n_x es el número de secuencias del alineamiento con el aminoácido *x* en una posición *j* y p_x es la frecuencia media de aparición del aminoácido *x* en todas las proteínas de la base de datos. Con lo cual podemos obtener un vector con las probabilidades binomiales de cada aminoácido *x* para una posición *j*.

$$\vec{P}_j=\left\{P_j^{\,ala},P_j^{\,cys},P_j^{\,asp},...,P_j^{\,tyr}\right\} \tag{46}$$

Este vector representa todos los cambios en la distribución de un aminoácido en relación a su conservación en una posición *j*, pero la probabilidad de que exista un aminoácido x en una posición *j* relativa a la probabilidad de que exista en otra posición *i* se puede relacionar con la energía libre estadística que separa las dos posiciones $\Delta G^x_{j\to i}$ (en una distribución de Boltzmann) mediante la siguiente expresión:

$$\frac{P_j^x}{P_i^x}=e^{\frac{\Delta G^x_{j\to i}}{kT^*}} \tag{47}$$

Donde kT* es una unidad arbitraria de energía y P_j^x es la probabilidad de que exista un aminoácido x en una posición *j*.

Si en la ecuación (47) consideramos *i* como una posición hipotética donde todos los aminoácidos se encontraran en su frecuencia media del alineamiento, de manera que sirva como estado de referencia para todas las posiciones, pudiendo considerar P_i^x como la frecuencia de aparición media de un aminoácido *x* en todo el alineamiento de secuencias (llamaremos a este término P_{Al}^x), nos será posible transformar $\overrightarrow{P_j}$ en un vector de energías estadísticas mediante la siguiente ecuación:

$$\Delta G_j^x = kT^* \ln\left(\frac{P_j^x}{P_{Al}^x}\right) \tag{48}$$

Donde ΔG_j^x es la energía libre estadística de cada aminoácido *x* en una posición *j*. Dicho vector podría definirse con el parámetro ΔG_j^{stat}, su energía estadística total:

$$\Delta G_j^{stat} = kT^* \sqrt{\sum_x \left(\ln \frac{P_j^x}{P_{Al}^x}\right)^2} \tag{49}$$

Finalmente para medir el acoplamiento existente entre posiciones, se calculará la energía del vector de una posición *j* en dos condiciones:

- **a.** En el alineamiento de secuencias completo: ΔG_j^{stat}
- **b.** En el subconjunto de secuencias, que representan la perturbación en una posición *i*, $\Delta G_{j|\delta i}^{stat}$.

De forma que la diferencia de energía de los dos vectores $\Delta\Delta G_{j,i}^{stat}$, dada por la ecuación (50), será la energía de acoplamiento estadística entre las dos posiciones *j* e *i*.

$$\Delta\Delta G_{j,\,i}^{stat} = kT^* \sqrt{\sum_x \left(\ln \frac{P_{j|\delta i}^x}{P_{Al|\delta i}^x} - \ln \frac{P_j^x}{P_{Al}^x}\right)^2} \tag{50}$$

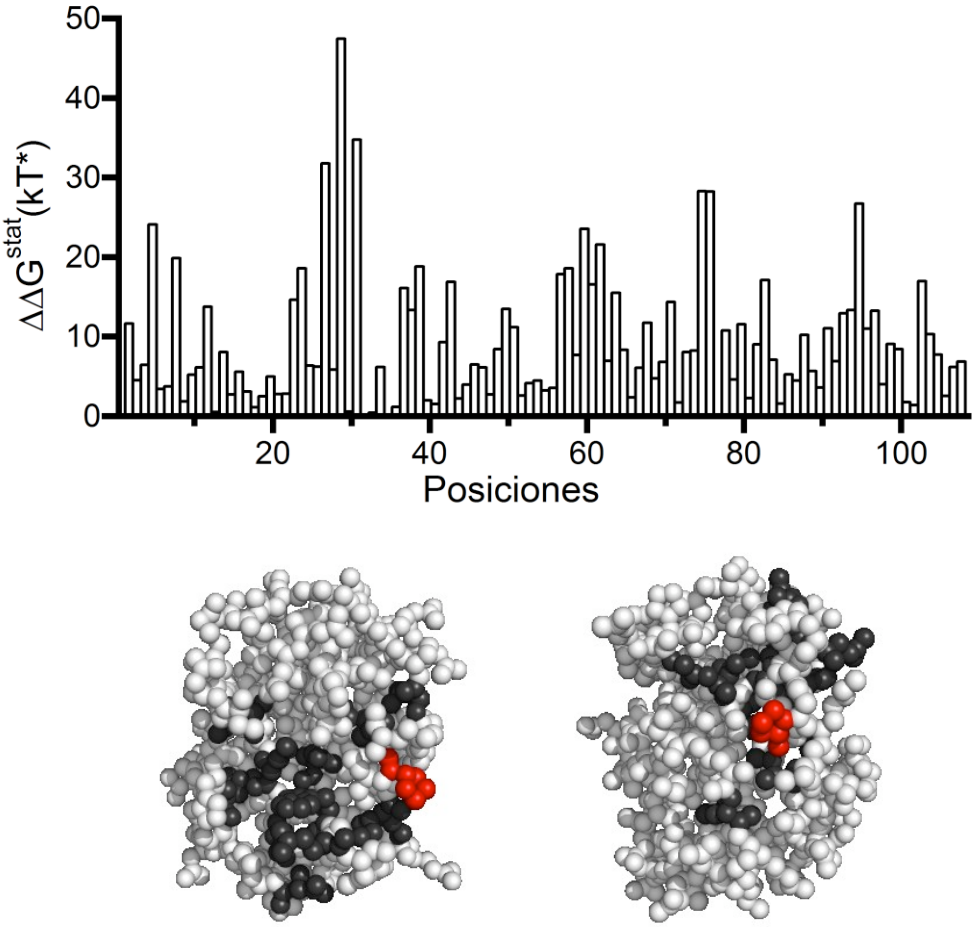

Figura 31. El gráfico de arriba corresponde a un análisis de acoplamiento estadístico realizado sobre un alineamiento de secuencias sobre la tiorredoxina de *E. coli* usando la base de datos de Uniprot/TrEMBL, el análisis esta hecho tomando como perturbación His en la posición 34, de las 1440 secuencias del alineamiento 132 tenían His en la posición 34. Abajo una representación de la estructura de tiorredoxina, el residuo en rojo corresponde a la posición 34 mientras que los residuos oscuros corresponden a las posiciones que mayor acoplamiento tienen con la 34, que aunque espaciadas en la secuencia están bastante próximas en la estructura terciaria como suele ser frecuente en el caso de las posiciones estadísticamente acopladas [93].

2.6.3.2 Análisis de covarianza simple

En este trabajo también se utilizó un modelo para el cálculo de la covarianza entre dos posiciones j e i [95], esta es calculada usando la siguiente expresión:

$$\sigma_{ji} = \sum_{\text{secuencias}} \frac{(\delta_j - \langle\delta_j\rangle)\cdot(\delta_i - \langle\delta_i\rangle)}{N_S} = \langle\delta_j\cdot\delta_i\rangle - \langle\delta_j\rangle\cdot\langle\delta_i\rangle \quad (51)$$

Donde N_S corresponde al número de secuencias que compongan el alineamiento, δ_p, siendo P igual a j o i, tomará valores de 1 o 0, respectivamente, si el residuo de la posición P de una secuencia es el mismo que el de la secuencia problema del alineamiento o diferente y $\langle\delta_p\rangle$ será el valor medio de δ_p en todas las secuencias del alineamiento ($\langle\delta_p\rangle = \frac{\sum \delta_p}{N_S}$)

2.7 Diseño de interacciones electrostáticas de la superficie de una proteína.

Una de las estrategias más recurrentes para la estabilización proteínas es la del diseño de distribuciones de cargas de la superficie que sean más favorables que la distribución de partida. Para ello el primer paso será necesariamente estimar la energía de interacción entre las cargas de una proteína, para de esta forma poder discernir que distribuciones de cargas son favorables, cual no lo son y en qué medida. Dicho cálculo se realizó usando una implementación desarrollada en nuestro grupo [96, 97] del modelo de Tanford y Kirkwood [98-100], en este modelo, ver Figura 32, una proteína es considerada como una esfera de baja constate dieléctrica, y los grupos cargados de la proteína como puntos fijos en dicha esfera; de manera que las interacciones electrostáticas entre ellos son las únicas interacciones tenidas en cuenta.

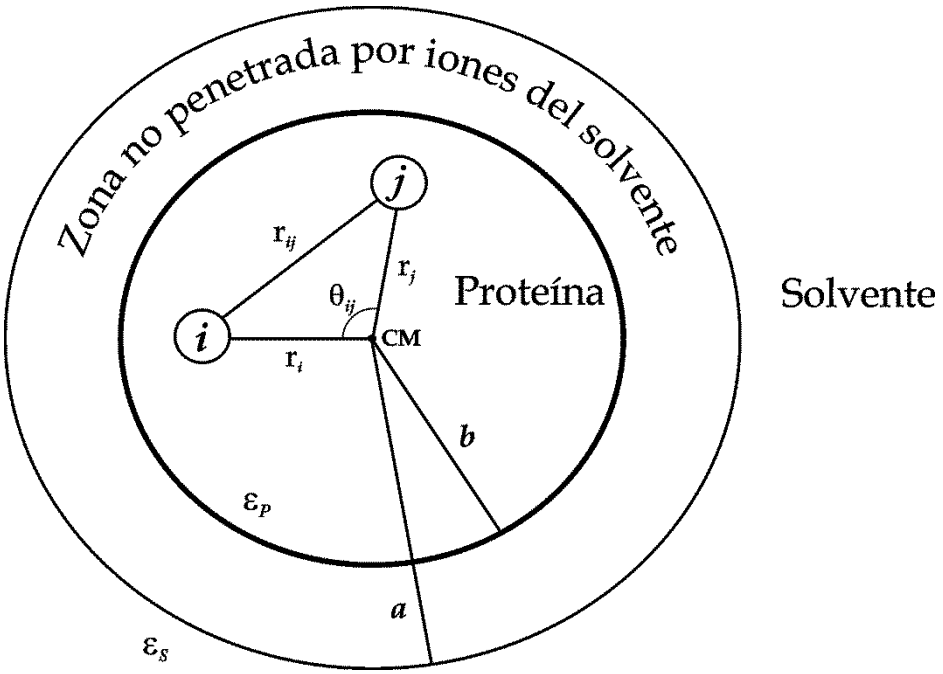

Figura 32. Esquema del modelo de Tanford y Kirkwood para el cálculo de interacciones entre residuos cargados. La proteína es considerada como una esfera de radio b, con una constante dieléctrica ε_P, la proteína esta englobada en una esfera mayor de radio a en la que no pueden penetrar los iones del solvente. Rodeando a ambas esferas se encuentra el solvente con una constante dieléctrica, ε_S, alta. Las cargas están representadas por las esferas *i* e *j*, donde r_j y r_i son la distancia de estas al centro de masas CM, y r_{ij} es la distancia entre ellas.

La energía de interacción entre dos residuos cargados de la superficie *i* y *j* vendrá dada por la siguiente expresión:

$$E_{ij}=e^2\left(\frac{A_{ij}-B_{ij}}{2b}-\frac{C_{ij}}{2a}\right)\left(1-SA_{ij}\right) \qquad (52)$$

Donde e es la unidad de carga, *b* el radio de la proteína que está relacionado con el volumen de la proteína (que se calcula a partir de la masa molecular de la proteína usando el valor típico para el volumen específico: 0.72mL/g), *a* se considera igual a *b* + 1.4 Å que es la media de los radios de catión y anión para el NaCl, A_{ij}, B_{ij} y C_{ij} fueron definidas previamente por Tanford y Kirkwood [98]: A_{ij} es la energía de interacción entre las cargas i y j en el interior de la proteína y es función de la constante dieléctrica de la proteína (fue usado el valor ε_P=4) y de la distancia entre las cargas r_{ij}; B_{ij} manifiesta la

contribución de las constantes dieléctricas de la proteína y solvente (para el solvente fue usada la constante dieléctrica del agua ε_S=78.5), así como las coordenadas de las cargas definidas por r_i, r_j y θ_{ij} y C_{ij} es función de las posiciones de las cargas y de la fuerza iónica del solvente. Por último SA_{ij} es la media de la accesibilidad al solvente (ASA) de los grupos *i* y *j*.

Para el cálculo de las fracciones de protonación de los grupos ionizables a un determinado pH y el de la energía asociada a toda una distribución de cargas de la proteína se utilizó el modelo de campo medio de Tanford y Roxby [101].

2.7.1 Diseño de distribuciones de carga estabilizantes

Aunque puede haber muchas posibilidades de mejorar las interacciones electrostáticas de las cargas superficiales de una proteína, muchas de las cuales supondrían la sustitución de uno o unos pocos residuos, lo ideal sería poder encontrar la distribución o distribuciones que resulten más favorables para una determinada secuencia y estructura. El problema es que esto supondría realizar una búsqueda exhaustiva entre todas las posibles distribuciones que pudiera tener una proteína, con lo cual para n residuos superficiales susceptibles de ser cargados, pudiendo estos tener carga +1, -1 o 0, tendríamos 3^n posibles distribuciones, el cálculo de la energía de cada una de ellas, aun en el caso de tratarse de proteínas pequeñas, estaría completamente fuera de nuestro alcance por requerir tiempos de computación inasumibles hoy en día. Otro problema añadido en este tipo de cálculos es que es probable encontrar un paisaje de energía rugoso, en el que haya una distribución óptima de cargas, pero posiblemente con muchos mínimos locales de similar energía entre ellos (ver Figura 33). Una manera eficaz de abordar estos problemas es el uso de un algoritmo genético [23, 102-104], capaz buscar en el espacio de distribuciones de cargas, pudiendo encontrar soluciones estabilizantes, simplificando mucho la búsqueda y por tanto el volumen de cálculos necesarios.

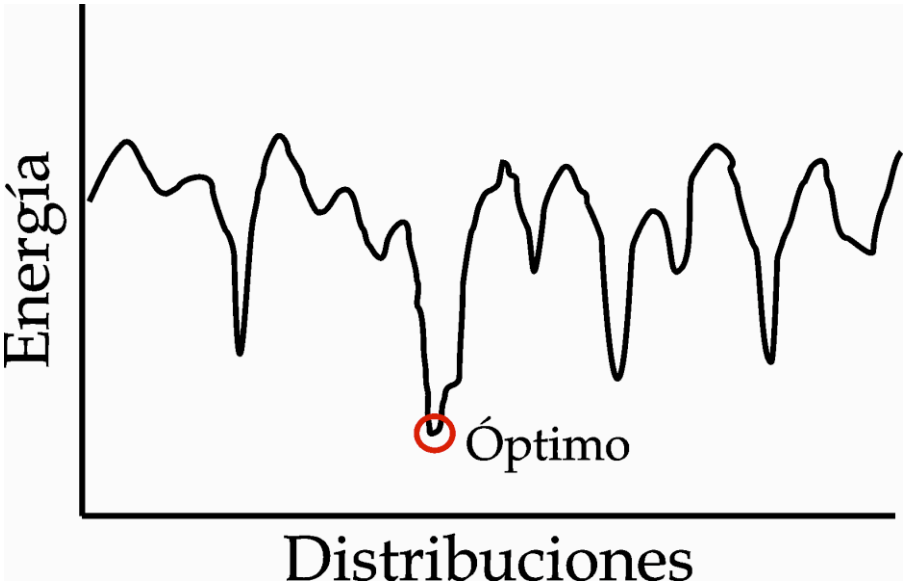

Figura 33. Ejemplo de "paisaje" de energía rugoso, con un óptimo y varios mínimos locales de similares energías.

Los algoritmos genéticos están inspirados en la teoría de la evolución biológica. En ellos, se parte de una población de soluciones, en nuestro caso de distribuciones de carga, a las cuales se las conoce como cromosomas, y se los aplica un proceso análogo a la evolución biológica, sometiéndolos a recombinaciones y mutaciones aleatorias, para posteriormente sufrir una selección en base a una determinada propiedad o "score". Para una descripción más detallada sobre los algoritmos genéticos, se puede consultar la siguiente web: http://www.obitko.com/tutorials/genetic-algorithms

2.7.2 Descripción del algoritmo genético utilizado

El algoritmo genético utilizado fue desarrollado en nuestro grupo de investigación [23]. Antes de iniciar el algoritmo genético se seleccionará un conjunto de n residuos, serán escogidos aquellos que tengan un porcentaje de área accesible al solvente mayor del 50%. Cada uno de estos residuos será susceptible de mutarse a carga positiva, negativa o neutra para conformar las distintas distribuciones de cargas posibles en la proteína. Estas distribuciones de carga son representadas en la Figura 34 como una ristra de cargas a las que

damos el nombre de cromosomas (ej: `++--0+--++00-`) donde `+` corresponderá a residuos cargados positivamente, `-` a residuos cargados negativamente y `0` a residuos neutros. El algoritmo dispondrá una primera generación de cromosomas aleatorios, cada uno de los cuales tendrá asociado un "score" Z, que será:

$$Z=E_{TK}+\delta \cdot N_{MUT} \quad (53)$$

Donde E_{TK} es la energía de interacción carga-carga para una distribución dada calculada según el modelo de Tanford y Kirkwood, δ será una penalización por mutación y N_{MUT} el número de mutaciones del cromosoma respecto a la proteína de partida, el sumando de la derecha que se puede incluir con objeto de obtener cromosomas que no tengan un excesivo número de mutaciones respecto a la proteína de partida, pero en muchos casos, como es el de este trabajo se prescindió de él.

El mejor cromosoma de esta primera generación, aquel que tenga el Z más bajo, pasará a la siguiente generación sin modificación alguna, los demás podrán ser seleccionados en parejas con una probabilidad proporcional a su "score" para sufrir un sobrecruzamiento, y dar lugar a su vez a dos nuevos cromosomas "hijos", por tanto la probabilidad de un cromosoma de ser elegido para sobrecruzamiento será proporcional a Z_W-Z, donde Z es el score de dicho cromosoma y Z_w es el score del peor cromosoma de la generación. El punto de sobrecruzamiento entre los cromosomas "padres" es aleatorio. Tras esto se somete a la nueva población de cromosomas a mutaciones puntuales, el programa nos permitirá fijar la probabilidad de una determinada posición a ser mutada, así los cromosomas resultantes junto al mejor cromosoma de la primera generación, que pasa sin modificarse, conformarán la segunda generación, por último se calculan los scores de esta segunda generación y volverá a empezar el ciclo. El proceso continuará hasta que se encuentre un

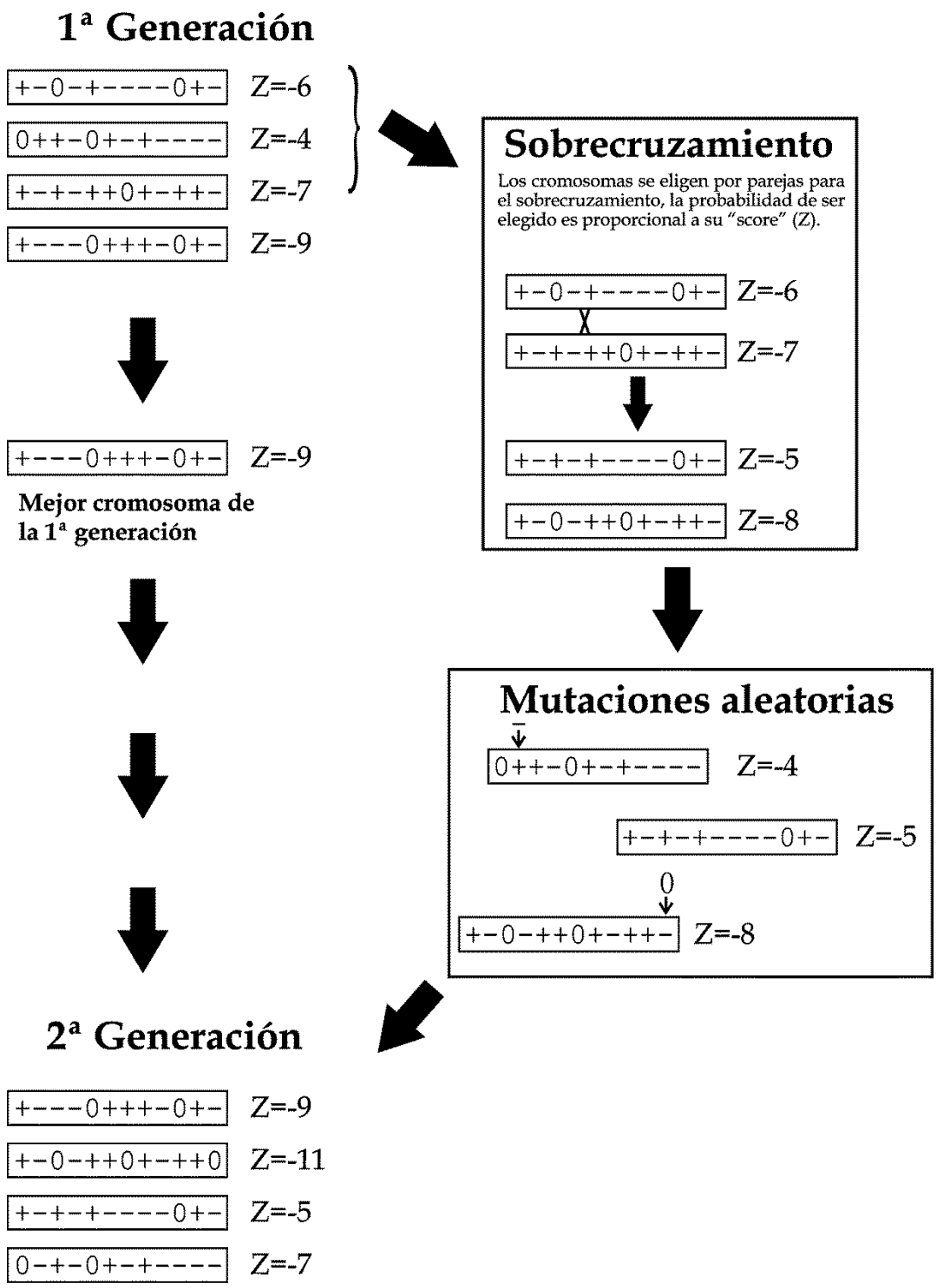

Figura 34. Esquema del funcionamiento del algoritmo genético desarrollado por nuestro grupo.

"mejor" cromosoma permanezca invariable durante un número determinado de ciclos, dicho número de ciclos nos permite fijarlos el programa. Se considerará que un cromosoma corresponde a un mínimo local si éste no puede ser mejorado por ninguna mutación puntual. En el presente trabajo solo se consideraron los cromosomas que correspondían a mínimos locales.

En la Figura 34 se haya representado un esquema del proceso que se acaba de describir.

El algoritmo genético utilizado en este trabajo, ha ce uso de una serie de parámetros importantes:

- Probabilidad de sobrecruzamiento: La idea es que cromosomas padres buenos, quizás puedan engendrar cromosomas hijos mejores, por tanto, fijamos la probabilidad de cada cromosoma a sufrir sobrecruzamiento haciéndola proporcional a su "score". Será Z_W-Z como sugerimos anteriormente.
- Probabilidad de mutación: Es la probabilidad de que se dé una mutación en una determinada posición del cromosoma (locus por analogía genética), si fijamos esa probabilidad al 100% todo el cromosoma será mutado, y por tanto, el algoritmo genético será equiparable a realizar una búsqueda aleatoria. En este trabajo fijamos la probabilidad de mutación en un 0.5% que es lo sugerido por el trabajo pionero de nuestro grupo [23].
- Tamaño de la población, o lo que es lo mismo, ¿cuántos cromosomas incluimos por generación?, cuantos menos cromosomas, menor será el espació de distribuciones de carga explorado, ya que existirán pocas posibilidades de sobrecruzamiento. Sin embargo, si la población es demasiado grande, el algoritmo genético puede ralentizarse demasiado. Nosotros establecimos una población de 20 cromosomas.

Una de las principales ventajas del algoritmo genético usado, es que éste puede encontrar soluciones en el espacio de distribuciones que cumplan una

serie de condiciones que pueden ser establecidas a priori, como por ejemplo, no incluir en los cálculos ciertos residuos funcionalmente importantes, o buscar distribuciones estabilizantes que tengan las mínimas diferencias posibles con la secuencia de partida.

3 Resultados

3.1 Artículo 1: “How many ionizable groups can sit on a protein hydrophobic core?

Es bien sabido que los aminoácidos ionizables cuando se encuentran formando parte de proteínas solubles suelen estar ubicados en la superficie, ya que pueden formar interacciones favorables con el agua, mientras que el interior de las mismas está prácticamente monopolizado por aminoácidos hidrófobos que rehúyen el contacto con el agua, y forman interacciones hidrófobas entre ellos. Esta distribución se considera clave no solo para que se produzca el plegamiento de la proteína, si no para dotarla también de cierta estabilidad que no comprometa el desempeño de su función.

Sin embargo, y a pesar del efecto altamente desestabilizante que se supone a la presencia de residuos ionizables en el ambiente hidrófobo que caracteriza al interior de una proteína, estos casos se presentan con mayor frecuencia de lo que cabría esperar, muchas veces fuertemente vinculados a la función (catálisis y otros procesos de transducción de energía) [105-109], y frecuentemente formando puentes salinos que podrían compensar el efecto desestabilizante que supone su ubicación [110]. Por tanto, resulta evidente que ser capaces de situar residuos ionizables en posiciones enterradas, aumentaría las posibilidades de la ingeniería de proteínas, haciendo más accesible, por ejemplo, el diseño de nuevas actividades catalíticas. ¿Pero cómo de resistentes son las proteínas a mutaciones de residuos hidrófobos por ionizables en su núcleo?

El objetivo de este trabajo es testar la tolerancia de una proteína a la simultánea introducción de varios residuos ionizables enterrados. Varios trabajos precedentes, como los llevados a cabo por el grupo de investigación de Bertrand García Moreno de la Universidad de Johns Hopkins [105, 106] o por nuestro propio grupo [111], habían sido capaces de ubicar un residuo ionizable en posiciones internas, hecho que conllevaba un notable descenso en la estabilidad de la proteína, pero no viéndose afectada su estructura y función. Muchos centros activos presentes en las proteínas no implican la presencia de un residuo ionizable enterrado, sino de varios, algo que a priori

parece que supondría esencialmente destruir el núcleo hidrófobo de la proteína. Para abordar el problema decidimos trabajar sobre una variante super-estable de la tiorredoxina [34]. Elegimos una serie de posiciones enterradas ocupadas por aminoácidos hidrófobos, y decidimos realizar una biblioteca combinatorial, de forma que cada posición elegida pueda ser mutada a Lisina (carga +), a Glutámico (carga -), o permanecer sin mutación (carga 0). El objeto de esta estrategia es buscar combinaciones favorables entre residuos cargados, y de esta forma atenuar el perjuicio sobre la estabilidad de la proteína. Las posiciones elegidas para formar la biblioteca se escogieron para que estuvieran razonablemente lejos del centro activo, y así aumentar la probabilidad de obtener variantes activas.

El estudio de un número de variantes de esta biblioteca combinatorial reveló un escenario muy sorprendente, ya que una buena proporción de las variantes estudiadas se plegaban, pero además, conservaban esencialmente la estructura y la capacidad enzimática de la proteína de partida. Quizás el resultado más esclarecedor fue la relación encontrada entre la estabilidad de las variantes respecto al número de mutaciones que acumulaban, revelando una caída de la estabilidad por mutación bastante suave, no drástica como cabría esperar, pudiendo llegar a obtener variantes que acumulaban hasta 4 mutaciones. Por otra parte la resistencia de la proteína a este tipo de mutaciones pareció ser independiente del signo de la carga de los residuos mutados, lo que hace pensar que la presencia de estos residuos pudo ser estabilizada más por un incremento en la flexibilidad conformacional de la proteína, que pueda permitir la exposición parcial de estos grupos al solvente, que por las interacciones favorables que pudieran surgir entre ellos.

Este trabajo sugiere en cualquier caso que las proteínas son mucho más robustas de lo esperado a la presencia de grupos ionizables enterrados, y parece que el único requisito importante para obtener proteínas con estos grupos es ser lo suficientemente estables. Además, hemos estimado por vez

primera el efecto que tienen estas mutaciones sobre la estabilidad de una proteína, lo cual consideramos que es de gran utilidad en el campo de la ingeniería de proteínas, por ejemplo para abordar el diseño de nuevos centros activos que involucren residuos ionizables enterrados. Por otra parte desde el punto de vista de la evolución de proteínas contribuye a engrosar la importancia del papel que se cree que tiene la estabilidad en la evolución de nuevas funciones, proteínas con altas estabilidades pueden ser más "evolucionables", es decir pueden acumular mayor número de mutaciones disruptivas, relacionadas normalmente, con el desarrollo de una nueva función.

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

SHORT COMMUNICATION

How many ionizable groups can sit on a protein hydrophobic core?

Hector Garcia-Seisdedos, Beatriz Ibarra-Molero, and Jose M. Sanchez-Ruiz*

Facultad de Ciencias, Departamento de Quimica Fisica, Universidad de Granada, 18071-Granada, Spain

ABSTRACT

Full or partial burial of ionizable groups in the hydrophobic interior of proteins underlies the large modulation in group properties (modified pK value, high nucleophilicity, enhanced capability of interaction with chemical moieties of the substrate, etc.) linked to biological function. Indeed, the few internal ionizable residues found in proteins are known to play important functional roles in catalysis and, in general, in energy transduction processes. However, ionizable-group burial is expected to be seriously disruptive and, it is important to note, most functional sites contain not just one, but several ionizable residues. Hence, the adaptations involved in the development of function in proteins (through *in vitro* engineering or during the course of natural evolution) are not fully understood. Here, we explore experimentally how proteins respond to the accumulation of hydrophobic-to-ionizable residue substitutions. For this purpose, we have constructed a combinatorial library targeting a hydrophobic cluster in a consensus-engineered, stabilized form of a small model protein. Contrary to naïve expectation, half of the variants randomly selected from the library are soluble, folded, and active, despite including up to four mutations. Furthermore, for these variants, the dependence of stability with the number of mutations is not synergistic and catastrophic, but smooth and approximately linear. Clearly, stabilized protein scaffolds may be robust enough to withstand many disruptive hydrophobic-to-ionizable residue mutations, even when they are introduced in the same region of the structure. These results should be relevant for protein engineering and may have implications for the understanding of the early evolution of enzymes.

Proteins 2012; 80:1–7.

Key words: protein stability; mutation effects; disruptive mutations; protein robustness; polar group burial; enzyme catalysis.

INTRODUCTION

Ionizable residues in proteins tend to be located at the surface, where they can establish favorable interactions with the aqueous solvent. Still, a small number of ionizable groups are often found in the hydrophobic interior of many proteins. This is perhaps not too surprising, since burial (full or partial) may place ionizable groups in local environments leading to the large modulation in properties (large pK shift, for instance) required for catalysis.[1,2] Indeed, internal ionizable residues are known to play important functional roles, not only in catalysis but also in ion transport, homeostasis, light-activated processes and, in general, energy transduction.[1–5] Certainly, ionizable-group burial is expected to be disruptive. Nevertheless, recent pioneering work by Garcia-Moreno and coworkers has demonstrated that many proteins can tolerate one hydrophobic-to-ionizable substitution.[3,4] That is, mutating a hydrophobic residue to, for instance, glutamate or lysine is clearly destabilizing, but the protein may remain folded and functional after mutation, provided that its stability prior to mutation was sufficiently high. Following this lead, we have shown that the pK value of an internal residue thus created can be modulated using rational protein engineering.[6] These results are highly relevant for our understanding of molecular evolution, as they support that, likely, special structural

Additional Supporting Information may be found in the online version of this article.
Grant sponsor: Spanish Ministry of Science and Innovation; Grant numbers: BIO2009-09562, CSD2009-00088; Grant sponsor: Junta de Andalucia; Grant number: CVI-1668.
*Correspondence to: Jose M. Sanchez-Ruiz, Facultad de Ciencias, Departamento de Quimica Fisica, Campus Fuentenueva s/n, 18071-Granada, Spain.
E-mail: sanchezr@ugr.es.
Received 21 January 2011; Revised 29 July 2011; Accepted 5 August 2011
Published online 30 August 2011 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.23166

adaptations are not required to make possible the presence of functionally important internal ionizable residues and, therefore, the development of function may be determined to a significant extent by protein stability,[3,4] an scenario which is fully consistent with the known proposal that stability promotes evolvability.[7] Furthermore, they are also relevant from the point of view of protein engineering, as clearly demonstrated by the recent design of a switchable Kemp eliminase on the basis of a single hydrophobic-to-ionizable mutation at the bottom of an hydrophobic cavity.[8]

However, there is one essential issue to consider in this context. The studies referred to above have demonstrated that proteins can tolerate ONE hydrophobic-to-ionizable mutation, while many functional sites contain SEVERAL ionizable and/or charged residues. Examples abound:[1] several carboxylic acids are considered essential catalytic groups in glycosidases, ribonucleases, aspartic proteases and hen-egg white lysozyme; two lysine residues are thought to be involved in the catalytic mechanism of acetoacetate decarboxylase; the catalytic triad in serine proteases includes two ionizable residues (histidine and aspartate); two arginines interact with the carboxylate groups of the substrate in the active site of aspartate aminotransferase; glutamate and histidine residues play essential roles in the catalytic mechanism of triosephosphate isomerases; the pK values of the Schiff's base with retinal and several aspartic acid residues change during the photocycle in bacteriorhodopsin, driving the functionally relevant proton transfers[9]; internal arginine and histidine residues have been proposed to play an essential role in preventing proton permeation through the water channel aquaporin.[10]

In view of the above, the two key questions would actually seem to be: "are special adaptations (other than high stability) required to allow the introduction of SEVERAL functionally important internal ionizable residues?" and "can a protein tolerate SEVERAL hydrophobic-to-ionizable residue substitutions?". It would appear that the answers to these related questions should be "yes" and "no", respectively. First of all, it is relevant here that the recent directed-evolution studies of Tokuriki and Tawfik,[11] aimed at increasing protein evolvability through buffering of destabilizing mutation effects, did not lead to a significant number of hydrophobic-to ionizable residue substitutions. These authors performed random drifts *in vivo* for proteins under chaperonin overexpression. They certainly found an increase in the number of mutations in core residues (as compared with the results of random drifts carried out without chaperonin overexpression). However, for glyceraldehydophosphate dehydrogenase and phosphotriesterase none of the core mutations reported as arising from drifts with chaperonin overexpression involved the replacement of hydrophobic residues with ionizable or charged residues (see the ~400 mutations reported in Supporting Information Table 8 of Tokuriki and Tawfik[11]). From a more general viewpoint, it is well-known that apolar residues in protein structures cluster together, reflecting the fact that hydrophobicity (loosely defined as the tendency of apolar moieties to avoid aqueous environments) is one of the main thermodynamic forces that drive folding. Introducing one hydrophobic-to-ionizable residue substitution will perturb one structurally essential hydrophobic cluster. However, introducing several such mutations at neighboring positions could be expected to essentially destroy the cluster and prevent the protein from folding properly.

Overall, the results and reasoning summarized above would seem to argue against the feasibility of introducing several hydrophobic-to-ionizable residue substitutions in the same region of a protein. Despite this expectation, we deemed of interest to perform an experimental test of the possibility of going beyond the single-mutation level in this context. For this purpose, we constructed a combinatorial library targeting a hydrophobic cluster in a consensus-engineered, stabilized form of *E. coli* thioredoxin[12] and found that, contrary to the naïve expectation, many of the variants randomly selected from the library are soluble, folded and active, despite including a comparatively large number of hydrophobic-to-ionizable residue substitutions.

MATERIALS AND METHODS

The combinatorial library of thioredoxin variants was constructed using gene—assembly mutagenesis,[13] as previously described.[12] For ease of purification, the variants had a His_6 tag attached to the amino-terminal (i.e., roughly opposite to the targeted hydrophobic cluster). The purification of these His_6-tagged variants was carried out as previously described.[12] All the 40 variants randomly selected for purification were soluble in the high-salt buffer used in the last step of the purification. However, significant precipitation was observed in many of them upon dialysis against the low-salt buffer (50 m*M* Hepes, pH 7) used in the calorimetric and spectroscopic experiments, with 50% of the variants precipitating completely. These insoluble variants, therefore, were not subjected to biophysical characterization. Reductase activity was measured using a turbidimetric assay of the thioredoxin catalyzed rate of reduction of insulin,[14] as previously described.[6]

Differential scanning calorimetry experiments were carried out in a capillary VP-DSC microcalorimeter (MicroCal, General Electric) as previously described.[12] Far-UV circular dichroism spectra were acquired using a Jasco (Tokyo, Japan) J-715 spectropolarimeter, as previously described.[6]

Native PAGE was performed as described by the manufacturer (Instruction Manual Mini-PROTEAN ® Tetra Cell, BIO-RAD), using a continuous buffer system. Pro-

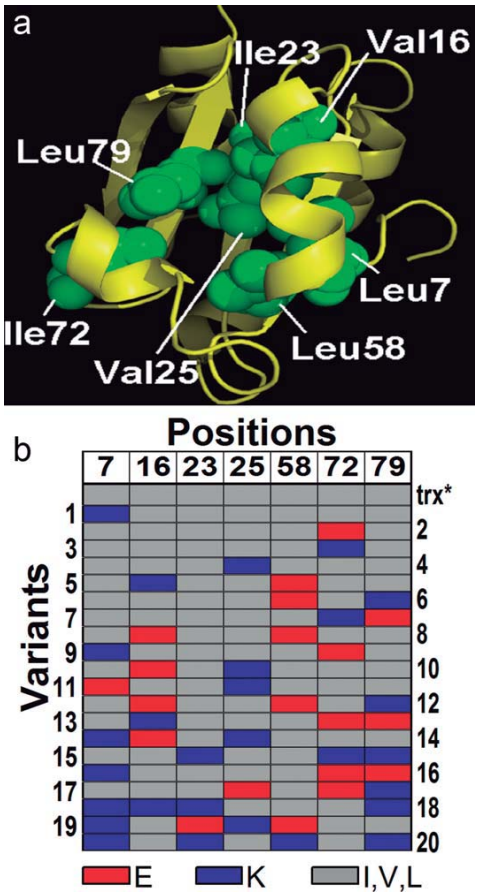

Figure 1

(**a**) Structure of thioredoxin (2trx) showing (green) the seven buried hydrophobic residues targeted for mutation. The side-chain accessible surface areas (in Å^2) for these residues are 0 (Leu7), 0 (Val16), 5.0 (Ile23), 0 (Val25), 0 (Leu58), 6.2 (Ile72), and 4.6 (Leu79). [Note that Ile75, also belonging to the same cluster, was not targeted because of its proximity to the active-site disulfide bridge]. A combinatorial library including all combinations of three possibilities (no mutation, mutation to Lys and mutation to Glu) over the seven positions was constructed. (**b**) Sequences of the library variants selected and analyzed in terms of structure, function and stability. Residues at the seven positions targeted are color-coded as gray (the hydrophobic side-chain present in the background trx* variant) red (Lys) and blue (Glu).

tein separation was carried out on 6% polyacrylamide gels. Totally 43 m*M* imidazole and 35 m*M* HEPES, pH 7.4 was used as continuous buffer according to McLellan.[15] Coomassie Blue staining was done following standard protocols. Final protein concentration was around 0.6 mg/mL. In order to assess the protonation state of our Lys/Glu trx* variants, a new set of protein variants was prepared to be used as charge electrophoretic markers. Thus, using trx* as background, a number of variants from a library targeting surface positions (manuscript in preparation) were selected for charge mutations. Since only solvent-exposed positions were mutated in these marker variants, there is little uncertainty regarding their total charge with respect to the trx* background. In particular, five different protein markers were obtained and purified as described previously.[12] Details regarding their sequence and extra charge taking trx* as reference are as follows: T14K/T54K/A105K (+3), H6E/T14K/T54K/Q62K/Q85K/Q98E/A105K (+3), T14K/T54K/N83D/Q85K/Q98E (+1), A68E/N83D/A105K (–1) and A50E/A68E/N83D/Q85K (−2). All 20 library variants [Fig. 1(b)] were studied by native PAGE (see Fig. S1 in Supporting Information for a representative example). To a good degree of approximation, integer values for the net charge (with respect to trx*) could be assigned to all the library variants through comparison with the markers (see Table I).

RESULTS AND DISCUSSION

In anticipation of the destabilizing character of the intended hydrophobic-to-ionizable residue mutations, we used as background a highly-stable (and fully active) variant of *E. coli* thioredoxin (trx*) previously obtained through consensus engineering.[12] We targeted 7 fully buried (accessible surface area ∼ 0) positions in the hydrophobic cluster involving the central β-sheet and helices 11–17 and 59–70 [Fig. 1(a)]. For each position we

Table I

Net Charges of the Thioredoxin Variants Described in Figure 1 and Possible Interpretations in Terms of Integer Charges at the Ionizable Residues

Variant	Net charge	Interpretations
Trx*	0	
1	1	K7+
2	−1	E72−
3	1	K72+
4	0	K25n
5	1	K16+ E58n
6	1	E58n K79+
7	0	K72+ E79− K72n E79n
8	−1	E16− E58n E16n E58−
9	0	K7+ E72− K7n E72n
10	−1	E16− K25n
11	0	E7− K25+ E7n K25n
12	0	E16− E58n K79+ E16n E58− K79+ E16n E58n K79n
13	−1	K16+ E72− E79− K16n E72− E79n K16n E72n E79−
14	1	K7+ E16− K25+ K7n E16n K25+ K7+ E16n K25n
15	3	K23+ K72+ K79+
16	−1	K7+ E72− E79− K7n E72− E79n K7n E72n E79−
17	−1	E25− E72− K79+ E25n E72− K79n E25− E72n K79n
18	3	K7+ K16+ K23+ K79n K7+ K16+ K23n K79+ K7+ K16n K23+ K79+ K7n K16+ K23+ K79+
19	1	K7+ E23n K25n E58n K7+ E23− K25+ E58n K7+ E23n K25+ E58− K7n E23n K25+ E58n
20	3	K7+ K23+ K58+ K79n K7+ K23+ K58n K79+ K7+ K23n K58+ K79+ K7n K23+ K58+ K79+

Note that values of the net charges are given with referente to trx* (i.e., net charge of variant minus net charge of trx*).

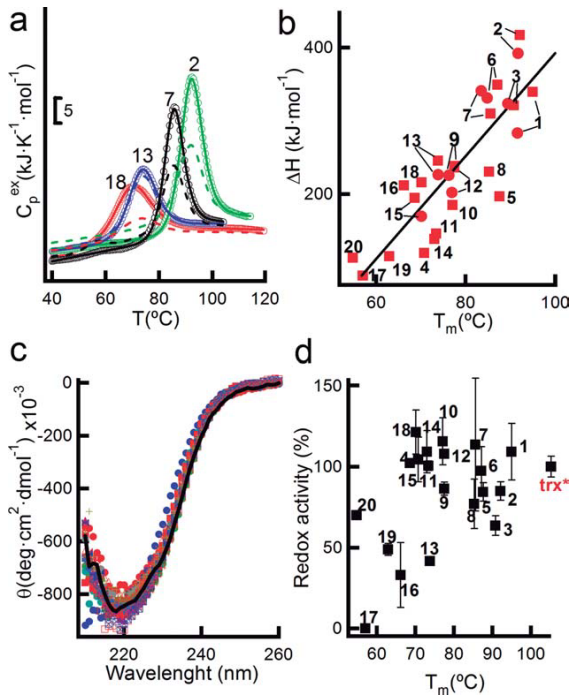

Figure 2

(**a**) Representative scanning calorimetry thermograms for the thermal denaturation of the thioredoxin variants studied in this work [sequences given in Fig. 1(b)]. Reheating runs are shown with dashed lines. Continuous lines are the best fits of a pseudo-two-state model to the first-heating data. (**b**) Plot of denaturation enthalpy versus denaturation temperature including all the studied thioredoxin variants. (**c**) Far-UV circular dichroism spectra for all the 20 studied thioredoxin variants. The spectra of the background trx* variant is shown with a continuous black line. (**d**) Plot of reductase activity (as percentage of the activity of the trx* background variant) versus denaturation temperature for all the 20 studied thioredoxin variants. Note that all variants (except variant number 17) display significant redox activity. The numbers in all the panels refer to the variants listed in Figure 1(b). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

considered three possibilities (no mutation, mutation to glutamate, mutation to lysine) and constructed a combinatorial library including all combinations of the 3 possibilities over the 7 positions. The library comprised 3^7 = 2187 variants of which 40 were randomly selected for purification. Of these, 20 variants [Fig. 1(b)] were soluble in low-salt buffer at the tenth of mg/mL level at least, and were amenable to characterization in terms of stability [Figs. 2(a,b) and 3(a)], structure [Fig. 2(c)] and function [Figs. 2(d) and 3(b)]. Examination of the sequences of the soluble [Fig. 1(b)] and low-solubility [Fig. 4(a)] variants reveals that a large number of mutations and the presence of a lysine at position 58 may contribute to the insolubility (see Fig. 4 for details). Circular dichroism spectra did not support drastic structural alterations in the soluble variants [Fig. 2(c)], which, in fact, retained in most cases significant reductase activity [Figs. 2(d) and 3(b)]. Scanning calorimetry showed well-defined thermal-denaturation transitions which displayed significant reversibility in most cases [Fig. 2(a)]. Calorimetric transitions were well described by pseudo-two-state model including a van't Hoff enthalpy to describe the temperature dependence of the unfolding equilibrium constant [Fig. 2(a)]. Van't Hoff enthalpy values were found in some cases to be higher than the corresponding calorimetric enthalpies. Van't Hoff to calorimetric enthalpy ratios larger than unity have been previously reported for *E. coli* thioredoxin and attributed to partial dimerization in both, the native and the unfolded state.[16] For a significant number of variants (13), the quality of the pretransition and post-transition baselines allowed reliable estimates of the unfolding heat capacity change to be

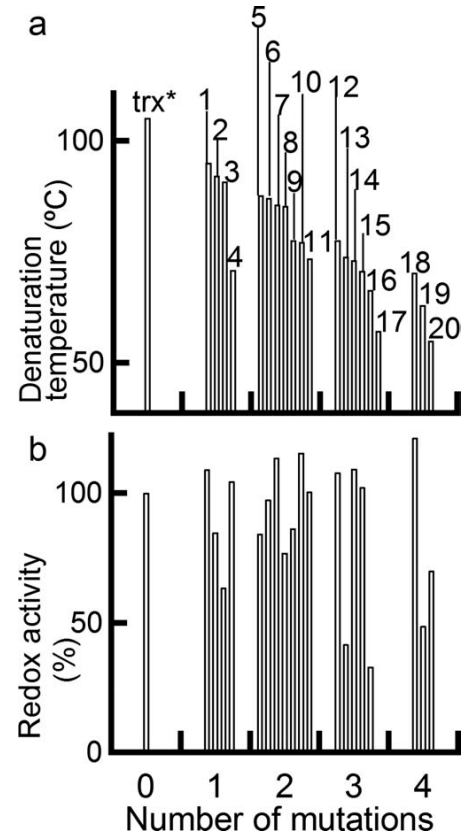

Figure 3

Denaturation temperature (**a**) and redox activity (**b**) of the thioredoxin variants specified in Figure 1(b) plotted versus the number of hydrophobic-to-ionizable residue mutations. Note in (a) the approximately linear dependence of the denaturation temperature value with the number of mutations. Redox activity values in (b) are given as percentages of the activity of trx* background variant. Note that all variants (except 17) display significant redox activity.

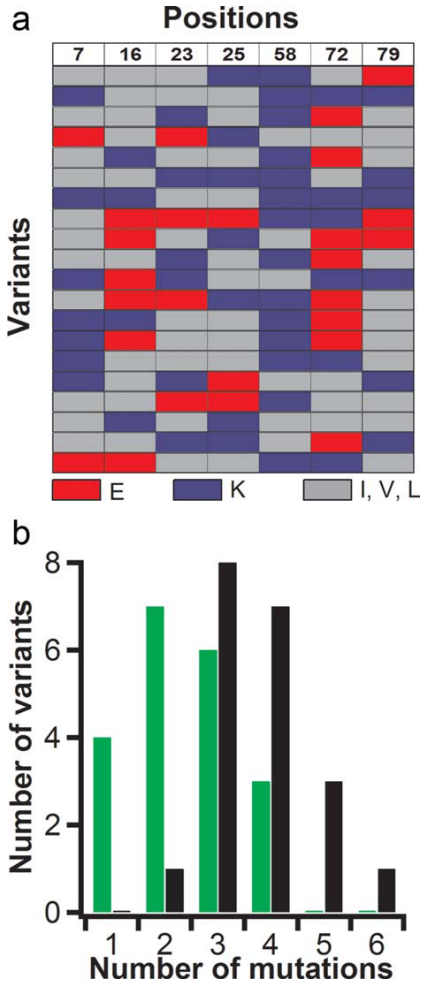

Figure 4

(**a**) Sequences of 20 thioredoxin variants that precipitated in the low-salt buffer used in the biophysical characterization studies. The color code is the same as that used in Figure 1(b) for the 20 variants that showed significant solubility in the low-salt buffer and were characterized in terms of structure, function and stability (Figs. 2 and 3). Note the presence of L58K in most of the insoluble variants. (**b**) Number of variants versus number of mutations distributions for the soluble variants [green, sequences given in Fig. 1(b)] and the insoluble variants (gray, sequences given in panel a of this figure). Note that insoluble variants tend to have a larger number of mutations. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

obtained from the experimental thermograms. These ΔC_P ranged between 4 and 7 kJ K^{-1} mol^{-1} with an average value of 5.1 $\pm$ 0.8 kJ K^{-1} mol^{-1}, in agreement with the experimental value for wild type thioredoxin unfolding: 5.8 $\pm$ 0.8 kJ K^{-1} mol^{-1}.[17] A plot of calorimetric enthalpy versus denaturation temperature [Fig. 2(b)] including all studied variants is roughly linear with a slope of 7.0 $\pm$ 0.8 kJ K^{-1} mol^{-1}. The fact that this slope is somewhat higher that the average ΔC_P value derived from the experimental calorimetric baselines, together with the dispersion observed in the plot of Figure 2(b), may be reflecting local structural reorganizations concomitant with the introduction of the ionizable groups at internal positions (see further below). Note, nevertheless, that the enthalpy scatter seen in Figure 2(b) has little influence on stability due to enthalpy-entropy compensation (see Fig. 5 and Fig. S2 and S3 in Supporting Information).

The most striking results were revealed by the values of the denaturation temperature, a distinct proxy of protein stability (see Fig. 5 for details on the relation between T_m and the unfolding free energy). As was to be expected, there is a stability price to be paid for the progressive destruction of a protein hydrophobic cluster. However, the dependence of stability on the number of hydrophobic-to-ionizable mutations [Fig. 3(a)] is not synergistic and catastrophic, but smooth and approximately linear, with a slope of about -10° per mutation, which, in terms of free energy is roughly equivalent to a decrease of about 7–8 kJ mol^{-1} per mutation (see Fig. 5). It is interesting to note in Figure 5 that the unfolding free energy at 37°C is only slightly above zero for the lowest stability variants. This suggests that the number of hydrophobic-to-ionizable substitutions that can be accepted is determined to a significant extent by the exhaustion of the stability at physiological temperature of the trx* variant used as background.

Finally, native PAGE was used to assess the total charge of the 20 variants studied. Table I collects the net charge values obtained as well as the possible interpretations in terms of integer charges on the mutated residues. While unique interpretations are not possible for most variants, it is clear that a significant number of the introduced ionizable residues must be charged in our library variants. To properly discuss this result, it is useful to review two different scenarios that emerge from published studies on buried ionizable groups in proteins. Statistical analyses of protein structures support that "wild-type" ionizable groups can sometimes be fully buried, provided that they form stabilizing hydrogen bonds and or salt bridges with other groups.[18,19] Introduction of hydrophobic-to-ionizable substitutions, on the other hand, likely leads to a different scenario, in particular when the ionizable groups become charged. For instance, recent molecular dynamics simulations on 18 variants of staphylococcal nuclease in which internal positions have been replaced by ionizable residues one at a time[20] suggest that backbone reorganization, localized partial unfolding, water penetration and increase in hydration occur concomitantly with residue charging. It is to be noted that this partial exposure of internal groups upon charging may actually be convenient from an enzyme engineering viewpoint, as the charged residue will be accessible to the substrate

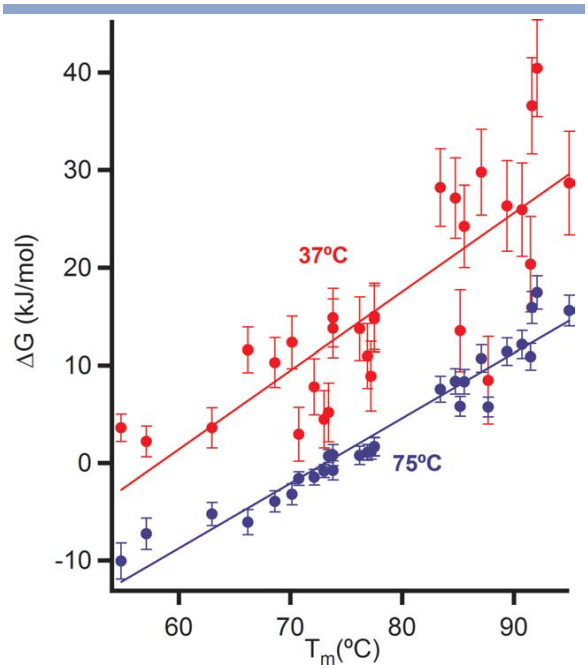

Figure 5

Correlation between unfolding free energies and denaturation temperature values for the variants characterized in this work (Figs. 1 and 2). ΔG values were calculated using the integrated Gibbs-Helmholtz equation:

$$\Delta G = \Delta H_m \cdot \left(1 - \frac{T}{T_m}\right) + \Delta C_p \cdot \left[T - T_m - T \cdot \ln\left(\frac{T}{T_m}\right)\right],$$

with the denaturation temperatures (T_m) and unfolding enthalpies at the T_m's determined from each calorimetric experiment [see Fig. 2(b) in the main text], together with an average value for the unfolding heat capacity change of 5.1 kJ K^{-1} mol^{-1} (see Fig. S2 in Supporting Information for a similar calculation using the actual ΔC_p values derived from the individual calorimetric experiments). This calculation assumes that the T_m value provides a good estimate of the temperature at which the free energies of the native and unfolded states are equal, even if there are deviations from two-state behavior (for details see Sanchez-Ruiz[24]). The ΔG values are calculated for 75 and 37°C, the noise being larger for the latter temperature reflecting the very long temperature-extrapolation. The estimated errors shown were calculated by the Monte Carlo method assuming Gaussian distributions for T_m, ΔH, and ΔC_p with standard deviations of 1.5°, 24.6 kJ mol^{-1} (from T_m and ΔH reproducibility in replicated experiments) and 0.76 kJ K^{-1} mol^{-1}, respectively. Straight lines are the best linear least squares fits and are meant to describe the general trends (actually, a small curvature may be seen in the ΔG vs. T_m plots). The slopes of these lines are 0.67 ± 0.03 kJ K^{-1} mol^{-1} (75°C) and 0.85 ± 0.11 kJ K^{-1} mol^{-1} (37°C). Accordingly, a decrease of 10° in denaturation temperature translates into a decrease of 7–8 kJ mol^{-1} in unfolding free energy. The ΔG values given in this figure should be considered as estimates, as they have not been corrected for potential kinetic distortions associated to partial irreversibility in the denaturation processes. Nevertheless, the ΔG vs. T_m correlations shown clearly support the use of the denaturation temperature as a proxy of thermodynamic stability in this case. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

but will still be in a microenvironment that leads to the modulation in properties useful in catalysis. For instance, NMR structure determination suggests that the accessibility to solvent of the catalytic Glu92 is roughly 10% in the Kemp eliminase[8] recently designed by DeGrado and coworkers (Ivan Korendovych and William DeGrado, personal communication); still, the presence of Glu92 in a hydrophobic cavity could enhance its basicity and facilitate the dehydration associated with substrate binding.

Overall, the existence of the two scenarios we have described above (fully buried charged "wild-type" groups stabilized by hydrogen bonds and/or salt bridges versus partial exposure to water upon charging of the residues resulting from a hydrophobic-to-ionizable substitution at an internal position) is consistent with the notion that the protein moeity, being rigid, and not polarizable, can stabilize internal charges in particular locations but is not likely to stabilize randomly introduced internal charges.[21] The relevant questions here are which scenario applies to the stabilization of the charges newly introduced (Table I) and how that stabilization relates to the unexpected success in preparing variants with a significant number of hydrophobic to-ionizable residue substitutions. Simple structural modeling (see Fig. S4 and S5 in Supporting Information) suggests that the thioredoxin scaffold used could accommodate a significant number of the fully buried ionizable groups in their neutral forms, at least in single-mutant variants. However, no specific design for stabilization of buried charges is involved in our multiple-mutant variants. Therefore, it appears likely that the charges created (Table I) are stabilized by increased-conformational-flexibility/partial-exposure mechanism. Certainly, we cannot rule out that favorable interactions between the polar groups of the glutamate and lysine residues (hydrogen-bonding, opposite charge interactions) may also contribute to avoid a precipitous drop in stability with the number of mutations. Note, however, that variants 18 and 20 [Fig. 1(b)] with 4 mutations to lysine, have denaturation temperature values [Fig. 3(a)] similar to that of variant 19, which has two mutations to glutamate and two mutations to lysine. Finally, it is also plausible that the methylene moieties of the lysine and glutamate side-chains could still participate in some kind of stabilizing hydrophobic interaction and thus contribute to the stabilization of our multiple mutant variants.

CONCLUSIONS

Although it must be recognized that the results reported here are to some extent open for molecular interpretation (see last paragraph in the preceding section), it is highly relevant that the stability/number-of-mutations dependence is nearly linear and, consequently, that variants with a large number of hydrophobic-to-ionizable residue substitutions can be readily prepared. In fact, variants with four mutations can be prepared,

although significantly less stable than the trx* background, show denaturation temperatures similar to those of many proteins from mesophilic organisms. It is likely that variants with even larger number of mutations could be obtained by using a further stabilized thioredoxin scaffold as background. Our results, therefore, have obvious implications for protein engineering and evolution that we briefly summarize below.

Replacing an internal hydrophobic residue with an ionizable one may lead to a polar group with the anomalous properties useful in enzyme catalysis (modified pK value, high nucleophilicity, enhanced capability of interaction with chemical moieties of the substrate, etc.) and several such internal ionizable residues may be required for biological function. We have shown that proteins may be robust enough to withstand many disruptive hydrophobic-to-ionizable residue mutations, even in the same region of the structure. Furthermore, we have provided a first experimental estimate of the associated pattern of stability penalties, which may be useful when selecting stabilized scaffolds for protein engineering tasks, such as the design of novel activities. Finally, we may speculate that the capability of high-stability proteins to accumulate many disruptive mutations may have played a role in the development of early enzyme catalysts. It is relevant in this context that ancestral reconstruction has revealed large stability enhancements for the oldest proteins.[22,23]

ACKNOWLEDGMENTS

The authors thank Ivan Korendovych and William DeGrado for useful comments and for sharing unpublished results.

REFERENCES

1. Harris TK, Turner GJ. Structural basis of perturbed pK_a values of catalytic groups in enzyme active sites. IUBMB Life 2002;53:85–98.
2. Warshel A, Sharma PK, Kato M, Parson WW. Modeling electrostatic effects in proteins. Biochim Biophys Acta 2006;1764:1647–1676.
3. Isom DG, Cannon BR, Castañeda CA, Robinson A, Garcia-Moreno BE. High tolerante for ionizable residues in the hydrophobic interior of proteins. Proc Natl Acad Sci USA 2008;105:17784–17788.
4. Isom DG, Castañeda CA, Cannon BR, Velu PD, Garcia-Moreno BE. Charges in the hydrophobic interior of proteins. Proc Natl Acad Sci USA 2010;107:16096–16100.
5. Karp DA, Stahley MR, Garcia-Moreno BE. Conformational consequences of ionization of Lys, Asp, and Glu buried at position 66 in staphylococcal nuclease. Biochemistry 2010;49:4138–4146.
6. Pey AL, Rodríguez-Larrea D, Gavira JA, Garcia-Moreno B, Sanchez-Ruiz JM. Modulation of buried ionizable groups in proteins with engineered surface charge. J Am Chem Soc 2010;132:1218–1219.
7. Bloom JD, Labthavikul ST, Otey CR, Arnold FH. Protein stability promotes evolvability. Proc Natl Acad Sci USA 2006;103:5869–5874.
8. Korendovych IV, Kulp DW, Wu Y, Cheng H, Roder H, DeGrado WF. Design of a switchable eliminase. Proc Natl Acad Sci USA 2011;108:6823–6827.
9. Luecke H, Richter H-T, Lanyi JK. Proton transfer in bacteriorhodopsin at 2.3 ansgstrom resolution. Science 1998;280:1934–1937.
10. Agre P. The aquaporin water channels. Proc Am Thorac Soc 2006;3:5–13.
11. Tokuriki N, Tawfik DS. Chaperonin overexpression promotes genetic variation and enzyme evolution. Nature 2009;459:668–673.
12. Rodríguez-Larrea D, Perez-Jimenez R, Sanchez-Romero I, Delgado-Delgado A, Fernandez JM, Sanchez-Ruiz JM. Role of conservative mutations in protein multi-property adaptation. Biochem J 2010;429:243–249.
13. Bessette PH, Mena MA, Nguyen AW, Daugherty PS. Construction of designed protein libraries using gene assembly mutagenesis. Methods Mol Biol 2003;231:29–37.
14. Holmgren A. Thioredoxin catalyzes the reduction of insulin disulfides by dithiothreitol and dihydrolipoamide. J Biol Chem 1979;254:9627–9632.
15. McLellan T. Electrophoresis buffers for polyacrylamide gels at various pH. Anal Biochem 1982;126:94–99.
16. Ladbury JE, Wynn R, Hellinga HW, Sturtevant JM. Stability of oxidized *Escherichia coli* thioredoxin and its dependence on the protonation of the aspartic acid residue in the 26 position. Biochemistry 1993;32:7526–7530.
17. Georgescu RE, Garcia-Mira MM, Tasayco ML, Sanchez-Ruiz JM. Heat capacity análisis of oxidized Escherichia coli thloredoxin fragments (1–73, 74–108) and their noncovalent complex. Evidence for the burial of apolar surface in protein unfolded states. Eur J Biochem 2001;268:1477–1485.
18. Rashin AA, Honig B. On the environment of ionizable groups in globular proteins. J Mol Biol 1984;173:515–531.
19. Bush J, Makhatadze GI. Statistical analysis of protein structures suggests that buried ionizable residues in proteins are hydrogen bonded or form salt bridges. Proteins 2011;79:2027–2032.
20. Damjauovic A, Brooks BR, Garcia-Moreno B. Conformational relaxation and water penetration coupled to ionization of internal groups in proteins. J Phys Chem A 2011;115:4042–4053.
21. Kim J, Mao J, Gunner MR. Are acidic and basic groups in buried proteins predicted to be ionized? J Mol Biol 2005;348:1283–1298.
22. Gaucher EA, Govindarajan S, Ganesh OK. Paleotemperature trend for Precambrian life inferred from resurrected proteins. Nature 2008;451:704–707.
23. Perez-Jimenez R, Ingles-Prieto A, Zhao ZM, Sanchez-Romero I, Alegre-Cebollada J, Kosuri P, Garcia-Manyes S, Kappock TJ, Tanokura M, Holmgren A, Sanchez-Ruiz JM, Gaucher EA, Fernandez JM. Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. Nat Struct Mol Biol 2011;18:592–596.
24. Sanchez-Ruiz JM. Ligand effects on protein thermodynamic stability. Biophys Chem 2007;126:43–49.
25. Sharp K. Entropy-enthalpy compensation: fact or artifact? Protein Sci 2001;10:661–667.

3.2 Artículo 2: "Probing the mutational interplay between primary and promiscuous protein functions: A computational-experimental approach"

Nuestro interés por la ingeniería de proteínas nos ha hecho fijar la atención en un fenómeno de creciente interés en los últimos años, la promiscuidad de proteínas. Así es como se conoce al hecho de que muchas de las enzimas hasta hoy caracterizadas posean otras funciones o actividades catalíticas aparte de su función nativa, para la que supuestamente han evolucionado. Estas actividades alternativas o promiscuas, a pesar de no realizarse con la misma eficiencia que la actividad nativa, rompen con la idea tradicional de "una enzima, un sustrato, una función" [112], revelando un escenario que reviste un enorme potencial biotecnológico [17], ya que el aumento de la eficiencia de funciones promiscuas nos llevaría a conseguir enzimas que actúen como catalizadores múltiples, siendo capaces de catalizar varios sustratos de forma eficaz.

Los estudios de evolución dirigida *in vitro* destinados a aumentar actividades promiscuas han observado el hecho de que a medida que se aumentan dichas actividades parece disminuir la actividad nativa de la proteínas, postulando la existencia de un trade-off entre dichas actividades [112].

En este trabajo nos planteamos abordar el reto de aumentar una actividad promiscua de una proteína usando la información contenida en los alineamientos de secuencias, comprobando de qué manera se vería afectada la actividad nativa, y así poder evaluar las directrices de la modulación simultánea de ambas. Para ello tomamos como proteína modelo la tiorredoxina de *E. coli*, que es una enzima reductasa, se encarga de reducir puentes disulfuro. Descrito está que una mutación en su centro activo, la mutación P34H, le confiere unos niveles apreciables de dos actividades promiscuas relacionadas con la asistencia para el correcto plegamiento de otras proteínas, como son la formación y reorganización de puentes disulfuro [64]. Con objeto de encontrar mutaciones que pudieran estar relacionadas con el aumento de estas actividades promiscuas, realizamos un análisis estadístico de acoplamiento (SCA) sobre un alineamiento de secuencias hecho

sobre tiorredoxina, para buscar las mutaciones que más correlacionen con P34H. Realizamos una pequeña biblioteca combinatorial con aquellas mutaciones de mayor acoplamiento. Para el estudio de la biblioteca combinatorial pusimos a punto un método novedoso que se compone del estudio experimental de un número reducido de variantes, la posterior predicción de toda la biblioteca combinatorial mediante la regresión PLS, y el uso de la frontera de Pareto, herramienta utilizada en economía para determinar las variantes óptimas para más de una propiedad, optimización multiobjetivo.

Para nuestra sorpresa, los resultados revelaron un simultáneo incremento de las actividades nativa y promiscuas de la tiorredoxina, lo cual supone la existencia de un patrón de modulación entre la actividad promiscua y nativa de una proteína que no había sido contemplado anteriormente [112]. Postulamos que este nuevo patrón de modulación no es incompatible con la existencia de trade-offs entre dos actividades, los resultados obtenidos parecen revelar la existencia de trade-offs que se encuentran definidos por la frontera de Pareto, es decir por las variantes optimas dentro del espacio combinatorial, mientras que la tiorredoxina silvestre, así como la inmensa mayoría de sus variantes, parecen encontrarse sub-optimizadas no sólo para su actividad promiscua, sino incluso para su actividad nativa, y por tanto son susceptibles de poder ser aumentadas por ambas actividades simultáneamente.

Estos resultados son, consecuentemente, de gran importacia en el campo de la ingeniería de proteínas. Está descrito que gran parte de las enzimas caracterizadas hasta el momento no están optimizadas en cuanto a su función nativa [113], de manera que a la vista de nuestros resultados se podría sugerir que es posible que muchas de ellas sean buenos puntos de partida para una optimización múltiple de varias actividades. Por otra parte la metodología experimental/computacional usada para el estudio de biblioteca

combinatorial desarrollado con éxito en este trabajo puede ser una herramienta muy eficaz en biotecnología, para el estudio de grandes bibliotecas combinatoriales de mutantes, cuando lo que buscamos es una optimización simultánea de varias propiedades.

OPEN ACCESS Freely available online

Probing the Mutational Interplay between Primary and Promiscuous Protein Functions: A Computational-Experimental Approach

Hector Garcia-Seisdedos, Beatriz Ibarra-Molero, Jose M. Sanchez-Ruiz*

Facultad de Ciencias, Departamento de Quimica Fisica, Universidad de Granada, Granada, Spain

Abstract

Protein promiscuity is of considerable interest due its role in adaptive metabolic plasticity, its fundamental connection with molecular evolution and also because of its biotechnological applications. Current views on the relation between primary and promiscuous protein activities stem largely from laboratory evolution experiments aimed at increasing promiscuous activity levels. Here, on the other hand, we attempt to assess the main features of the simultaneous modulation of the primary and promiscuous functions during the course of natural evolution. The computational/experimental approach we propose for this task involves the following steps: a function-targeted, statistical coupling analysis of evolutionary data is used to determine a set of positions likely linked to the recruitment of a promiscuous activity for a new function; a combinatorial library of mutations on this set of positions is prepared and screened for both, the primary and the promiscuous activities; a partial-least-squares reconstruction of the full combinatorial space is carried out; finally, an approximation to the Pareto set of variants with optimal primary/promiscuous activities is derived. Application of the approach to the emergence of folding catalysis in thioredoxin scaffolds reveals an unanticipated scenario: diverse patterns of primary/promiscuous activity modulation are possible, including a moderate (but likely significant in a biological context) simultaneous enhancement of both activities. We show that this scenario can be most simply explained on the basis of the conformational diversity hypothesis, although alternative interpretations cannot be ruled out. Overall, the results reported may help clarify the mechanisms of the evolution of new functions. From a different viewpoint, the partial-least-squares-reconstruction/Pareto-set-prediction approach we have introduced provides the computational basis for an efficient directed-evolution protocol aimed at the simultaneous enhancement of several protein features and should therefore open new possibilities in the engineering of multi-functional enzymes.

Citation: Garcia-Seisdedos H, Ibarra-Molero B, Sanchez-Ruiz JM (2012) Probing the Mutational Interplay between Primary and Promiscuous Protein Functions: A Computational-Experimental Approach. PLoS Comput Biol 8(6): e1002558. doi:10.1371/journal.pcbi.1002558

Editor: Arne Elofsson, Stockholm University, Sweden

Received December 28, 2011; **Accepted** April 29, 2012; **Published** June 14, 2012

Funding: This work was supported by FEDER Funds and Grants BIO2009-09562 and CSD2009-00088 from the Spanish Ministry of Science and Innovation. 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 declared that no competing interests exist.

* E-mail: sanchezr@ugr.es

Introduction

Proteins are capable to perform molecular tasks with impressive efficiency and, often, with exquisite specificity. Nevertheless, many proteins possess weak promiscuous functions, which are more or less related to the primary activity, but involve different substrates or different chemical alterations [1–5]. Protein promiscuity has been extensively studied in recent years due to its important biotechnological applications [6–12], to its role in adaptive metabolic plasticity [13–15] and also because of its fundamental connection with molecular evolution. Indeed, promiscuity in modern proteins is plausibly a vestige of the broad specificity of primordial proteins [1]. Furthermore, as briefly elaborated below, promiscuity likely plays an essential role in the development of new functions through divergent evolution [3,5,8,16–20].

Development of new functions does occur during evolution, sometimes with impressive speed. In most cases, the process involves gene duplication as a necessary step. It has been repeatedly noted, however, that random accumulation of mutations in a gene is unlikely to create a new function. It is generally assumed, therefore, that a sufficient level of the new (initially promiscuous) activity must be present before the duplication event. In this way, natural selection can act on one of the gene copies to enhance the new function, while the original function is retained by the other copy. However, optimization of a functional site for a given molecular task likely interferes with the efficient performance of the protein for a different task based on the same site. Consequently, enhancement of the promiscuous activity prior to gene duplication may be expected to cause a decrease in primary activity that could conceivably compromise organism survival. As a solution to this conundrum, a ''weak trade-off'' scenario has been proposed [5]: enhancement of the promiscuous activity is assumed to be accompanied with only a moderate decrease in primary function and, therefore, a generalist protein (significant levels of both activities) can be formed prior to gene duplication without seriously impairing organism fitness. This weak trade-off explanation is certainly supported by a number of laboratory evolution experiments [5]. However, the possibility that *natural* evolution may actually avoid or bypass primary/promiscuous activity trade-offs (i.e., a ''no trade-off'' scenario as opposed to a ''weak trade-off'' scenario) should be seriously taken into account, since bifunctional enzymes with the capability to catalyze efficiently

Author Summary

Interpretations of evolutionary processes at the molecular level have been determined to a significant extent by the concept of "trade-off", the idea that improving a given feature of a protein molecule by mutation will likely bring about deterioration in other features. For instance, if a protein is able to carry out two different molecular tasks based on the same functional site (competing tasks), optimization for one task could be naively expected to impair its performance for the other task. In this work, we report a computational/experimental approach to assess the potential patterns of modulation of two competing molecular tasks in the course of natural evolution. Contrary to the naïve expectation, we find that diverse modulation patterns are possible, including the simultaneous optimization of the two tasks. We show, however, that this simultaneous optimization is not in conflict with the trade-offs expected for two competing tasks: using the language of the theory of economic efficiency, trade-offs are realized in the Pareto set of optimal variants for the two tasks, while most protein variants do not belong to such Pareto set. That is, most protein variants are not Pareto-efficient and can potentially be improved in terms of several features.

two different biochemical reactions based on the same active site are known and have been recently characterized in detail [21,22] and experimental studies have supported that the trade-off between high activity and tight specificity can be greatly relaxed [23].

Here, we aim at assessing the patterns of primary/promiscuous activity modulation in the mutational space actually explored by natural evolution when recruiting a promiscuous activity for a new function. The approach we propose involves essentially three steps:

1) *A set of positions that are likely linked with the recruitment of the promiscuous activity for a new function is determined*. For this task we resort to statistical coupling analysis (SCA), a sequence-based bioinformatics procedure originally developed to determine networks of energetically-coupled, co-evolving residues in proteins [24]. SCA has been successfully applied to the design of allosteric communication in proteins [25] and to the design of artificial sequences able to fold to target structures [26].
2) *A combinatorial library of mutations on the set of positions determined in step 1 is prepared and assessed for both, the primary and the promiscuous activities*. In this regard, one methodological (but relevant) point must be made. Since high throughput screening is not generally available, we actually analyze a comparative small number of library variants and use a partial least-squares fitting to these data to reconstruct the whole combinatorial space. Partial least squares (PLS) [27] is a regression technique that allows the prediction of dependent variables (primary and promiscuous activities for the library variants, in the case of interest here) from a very large number of independent variables (related, in the case of interest here, to the individual mutation effects and to non-additivity –i.e., coupling- between mutation effects). PLS has been used in social sciences and chemometrics for many years and its application to protein design has been recently explored [28].
3) *An approximation to the Pareto set of variants of optimal primary/promiscuous activities is obtained from the library screening in order to assess the limits to the simultaneous enhancement of both activities*. The determination of the Pareto set is actually the key goal in the approach we propose. Therefore, we describe below in some detail the meaning of this concept and how it can be used to clarify the relation between the primary and promiscuous activities of a protein.

A unique global optimum cannot be defined when dealing with a multi-objective optimization problem, such as, for instance, enhancing a promiscuous activity while keeping the level of the primary activity as high as possible. However, a set of several optimal solutions can be defined using the Pareto criterion: a solution (protein variant in this case) belongs to the set of optimal solutions (the so-called Pareto set) if it is not *dominated* by any other solution. The dominance relationship is defined as follows: a solution **a** dominates a solution **b** if it shows enhanced performance for all optimization objectives. In the specific case of interest here, variant **a** dominates variant **b** if primary-activity(**a**)>primary-activity(**b**) and *simultaneously* promiscuous-activity(**a**)>promiscuous-activity(**b**). The construction of the Pareto set of non-dominated solutions is illustrated with a simple example in Figure 1A. The Pareto set includes the solutions with optimal trade-offs between the different objectives and has been used extensively in economics, while its application to protein design has only been explored in recent years [29–31]. In the specific case of interest here, determination of the Pareto set should immediately clarify the main features of the modulation of the primary and promiscuous activities within a given mutational space. For instance, if the starting variant (the "wild-type" protein, for instance) already belongs to the Pareto set, enhancement of the promiscuous activity necessarily implies a decrease in primary function and the trade-off will be weak (Figure 1B) or strong (Figure 1C) depending of the general slope of the Pareto set in the plot of promiscuous activity versus primary activity. On the other hand, if the starting variant does not belong to the Pareto set, simultaneous optimization of both activities is in principle feasible (Figure 1D) and primary/promiscuous trade-offs can be avoided to some extent.

To test the approach proposed, we have chosen the three basic activities associated with the thioredoxin fold: reduction of disulfide bridges, formation of disulfide bridges and isomerization (reshuffling) of disulfide bridges. The two latter activities are linked in vivo to protein folding processes [32,33] (oxidative folding and rescuing of proteins with incorrect disulfide bridges) and, in the periplasm of bacteria, are performed by different proteins: DsbA and DsbC, respectively. By contrast, in the endoplasmic reticulum of eukaryotic cells, both disulfide-linked folding processes are catalyzed by the same protein: protein disulfide isomerase (PDI). PDIs are multidomain proteins that contain thioredoxin-fold domains [32]. Processes of disulfide reduction in vivo (obviously unrelated with folding) are typically catalyzed by single-domain thioredoxins, which may also show low in vitro levels of the protein folding activities associated to disulfide bridge formation and reshuffling. Indeed, it is tempting to speculate that low levels of these activities were already present in primordial thioredoxins and that, at some evolutionary point, were recruited for new-function development leading to the proteins involved in disulfide-bridge-linked protein folding [32].

The processes of thioredoxin-domain catalyzed reduction, formation (oxidation) and reshuffling of disulfide bonds are all dependent on the active-site CXXC motif. Reduction (see Figure 2) starts with the reduced enzyme and involves the nucleophilic attack of the thiolate form of the amino-terminal cysteine on the disulfide bridge of the substrate [34,35]. The mixed-disulfide thus formed is resolved by the nucleophilic attack of the carbonyl-terminal cysteine. Oxidation, on the other hand

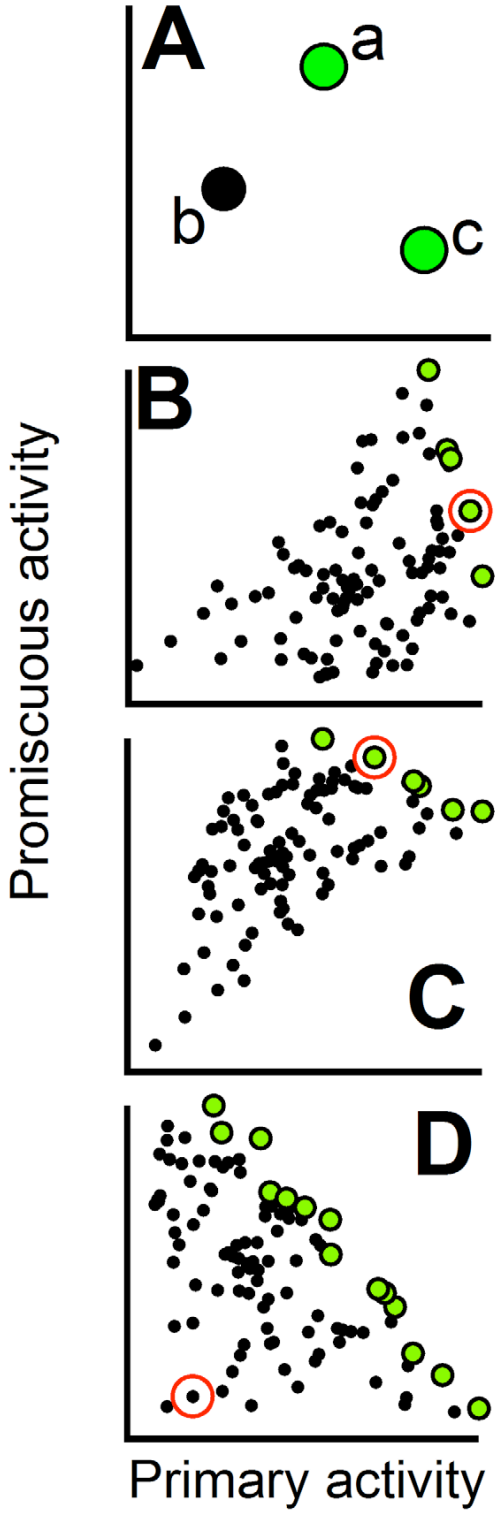

Figure 1. Use of the Pareto set to define the patterns of primary/promiscuous activity modulation. In all plots, each data point represents the primary and promiscuous activity data for a given protein variant. Different variants may be thought as corresponding to different combinations of mutations from a given set. (A) Illustration of Pareto set construction. Variant "b" is dominated by variant "a", since the latter shows higher values of *both* activities. Variant "a" is not dominated by variant "c", since promiscuous-activity(a)>promiscuous-activity(c). Variant "c" has the highest value for the primary activity and is not dominated neither by "a" nor "b". The non-dominated variants "a" and "c" form the Pareto set (green data points) for this three-variant example. (B), (C) and (D) are illustrative examples of the relation between the Pareto set (green data points) and the primary/promiscuous trade-offs. In (B) and (C) the starting variant (marked with a red circle) belongs to the Pareto set and, therefore, increasing the promiscuous activity necessarily implies a decrease of the primary activity. The plot in (B) is meant to illustrate a weak trade-off along the Pareto set (a significant increase in promiscuous activity can be achieved with only a small decrease in primary activity) while (C) is meant to illustrate a strong trade-off. In (D) the starting variant does *not* belong to the Pareto set and, hence, the simultaneous enhancement of both activities is possible.
doi:10.1371/journal.pcbi.1002558.g001

(see Figure 3), involves a nucleophilic attack of the substrate on the disulfide bridge of the oxidized enzyme and the mixed-disulfide intermediate is resolved by attack from the free cysteine (in the thiolate form) of the substrate [36]. It is relevant that the oxidation and reduction processes involve opposite chemical changes in the substrate (break-up and formation of disulfide bridges) as well as different mechanisms for the resolution of the mixed-disulfide intermediate. Furthermore, the two processes may be expected to be linked to different values of the redox potential (as suggested by the redox potentials of thioredoxin and PDI: see Figure 6 in Hatahet & Ruddock [36]) and, as it has been extensively discussed in the literature, they likely have different molecular requirements in terms of the conformational changes during catalysis, the stability of cysteine thiolates and the modulation of the pK values of the catalytic groups [33,36–39]. Clearly, disulfide reduction and catalysis of oxidative folding (involving formation of disulfide bridges) may be expected to strongly trade-off.

Contrary to disulfide reduction and formation (Figures 2 and 3), disulfide-bridge reshuffling in misfolded proteins to yield the correctly folded state does not involve a net change in the oxidation state of the substrate and could in principle occur through cycles of catalyzed reduction/oxidation [36,40]. Alternatively, the initial attack of the enzyme on a substrate disulfide bridge may yield a free cysteine that could attack another disulfide bridge thus starting a cascade of disulfide-bond rearrangments leading to the most stable configuration [36,40].

The specific protein system we use in this work is *E. coli* thioredoxin, an enzyme involved in multiple reduction processes in vivo [35,41] which, besides this primary (i.e. reductase) activity is able to catalyze, albeit with very low efficiency, disulfide-bridge-linked protein folding processes [42,43] (promiscuous activities of *E. coli* thioredoxin). We apply the approach proposed (steps 1–3 above) to *E. coli* thioredoxin with the catalysis of oxidative folding as the promiscuous activity. Nevertheless, the variants thus obtained are also tested for the disulfide reshuffling activity.

Results/Discussion

Statistical coupling analysis of the emergence of folding catalysis in the thioredoxin fold

To find a set of positions likely linked to the emergence of disulfide-bridge-linked folding functions in the thioredoxin fold, we have used statistical coupling analysis (SCA) which works by comparing the amino acid distributions at different positions in a multiple sequence alignment (MSA) with the corresponding ones in a given sub-alignment [24]. To apply SCA to the study of primary/promiscuous activity modulation we propose selecting

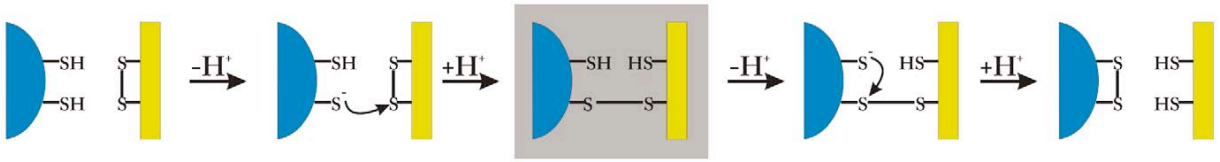

Figure 2. Basic mechanism of disulfide reduction catalyzed by a thioredoxin-fold domain. (Thioredoxin domain shown in blue). The mixed disulfide intermediate has been highlighted in gray. Note the resolution of this intermediate through attack of the thiolate form of the C-terminal cysteine group of the catalyst.
doi:10.1371/journal.pcbi.1002558.g002

the sub-alignment on the basis of a function-related criterion related with the promiscuous activity. We thus start with a MSA derived from a sequence-database search using the *E. coli* thioredoxin sequence as query and select as sub-aligment those sequences belonging to thioredoxin-fold domains in proteins involved in protein folding in vivo. Actually, this selection step is made straightforward by the fact that these domains contain the active-site CGHC sequence, while thioredoxin reductases contain the CGPC sequence. In fact, the P34H mutation on the *E. coli* thioredoxin background has been shown to enhance significantly its "PDI-like" promiscuous activities [43]. The MSA we have used contains indeed a significant number of sequences with a histidine at position 34 (*E. coli* thioredoxin numbering) that, in most cases, belong to thioredoxin-fold domains of eukaryotic PDI's. We thus obtained the statistical free-energies using the P→H substitution as the perturbation at position 34 (see Methods for details) and we expect these values to reveal networks of residues related with the emergence of the protein folding activities in the thioredoxin fold.

However, SCA is based upon the perturbation of the 20 amino acid distribution at each position and actually provides a list of coevolving positions (see Figure 4A), while we are interested in specific mutations at these positions. Therefore, we included an additional layer of statistical analysis. For each given position, we considered the mutation from the amino acid present in *E. coli* thioredoxin (the "Ec" aminoacid) to the amino acid "X" defined as the amino acid different from "Ec" that has the highest frequency (largest number of occurrences in the sequence alignment) when there is a histidine at position 34. Then, for each of the 13 positions with the highest coupling free energies from the SCA analysis (Figure 4A), we calculated the following score:

$$\Gamma_{Ec\to X} = \ln\left(f_X/f_{Ec}\right)_H - \ln\left(f_X/f_{Ec}\right)_P \qquad (1)$$

where f is the frequency of occurrence of the amino acids in the sequence alignments and subscripts "H" and "P" refer to the condition for the calculation of the frequencies (histidine or proline at position 34, respectively). Large positive values for the score indicate that the P34H substitution shifts the statistics strongly towards amino acid X. We retained for experimental analysis the 10 positions (and the corresponding Ec→X mutations) for which the score was positive (see Figure 4B): I4V, D26E, W28Y, E30P, I38L, K57A, N59D, D61T, I75Y, L94R. It is important to note that the 10 positions selected form a well-defined, connected network surrounding the active site (see Figure 4C), a fact fully consistent with their likely role in the development of the new functions related with protein folding catalysis. We also wish to emphasize at this point that the immediate purpose of this work (see section below) is to probe the interplay between primary and promiscuous activities in the mutational space defined by the 10 mutations selected. Specifically, the derivation of a molecular-level picture of what each of these mutations is doing (a task that would require extensive structural work due to the potential non-additivity of the mutation effects), is beyond the scope of this work.

Combinatorial library analysis and partial least-squares reconstruction of the full mutational space

A simple visual examination (see Figure 4D) of the sequences of the MSA used that include histidine at position 34 (most of them belonging to eukaryotic PDI's) shows that different combinations of the 10 mutations selected occur in extant PDI's. We conclude that natural selection does efficiently explore the mutational space defined by combinations of the 10 mutations. To assess how the interplay between the primary and promiscuous activities is modulated in this mutational space, we prepared a combinatorial library spanning the 10 mutations (i.e., $2^{10} = 1024$ variants) on the P34H background and determined the reductase and the catalysis of oxidative folding activities for 29 randomly selected variants. The primary activity has been probed by following the standard reduction assay for DTT-reduced thioredoxin which uses insulin as a model substrate (see Methods for details). Clearly, this assay reflects the reduction process of Figure 2. The catalysis of oxidative folding (promiscuous activity in *E. coli* thioredoxin) has been assessed using fully reduced ribonuclease A as substrate (see Methods for details). This assay might include some contribution from the reshuffling process since the first disulfide bridges formed need not be the correct ones. Nevertheless, it is expected to probe

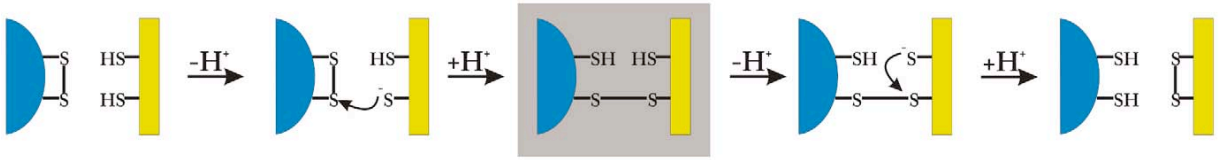

Figure 3. Basic mechanism for disulfide oxidation catalyzed by a thioredoxin-fold domain. (Thioredoxin domain shown in blue). The mixed disulfide intermediate has been highlighted in grey. Note that, unlike the mechanism of disulfide reduction shown in Figure 2, resolution of the intermediate occurs through attack of a thiolate group of the substrate. Actually, attack of the thiolate from the catalyst (as in Figure 2) must be prevented since it would revert the substrate to the initial reduced state.
doi:10.1371/journal.pcbi.1002558.g003

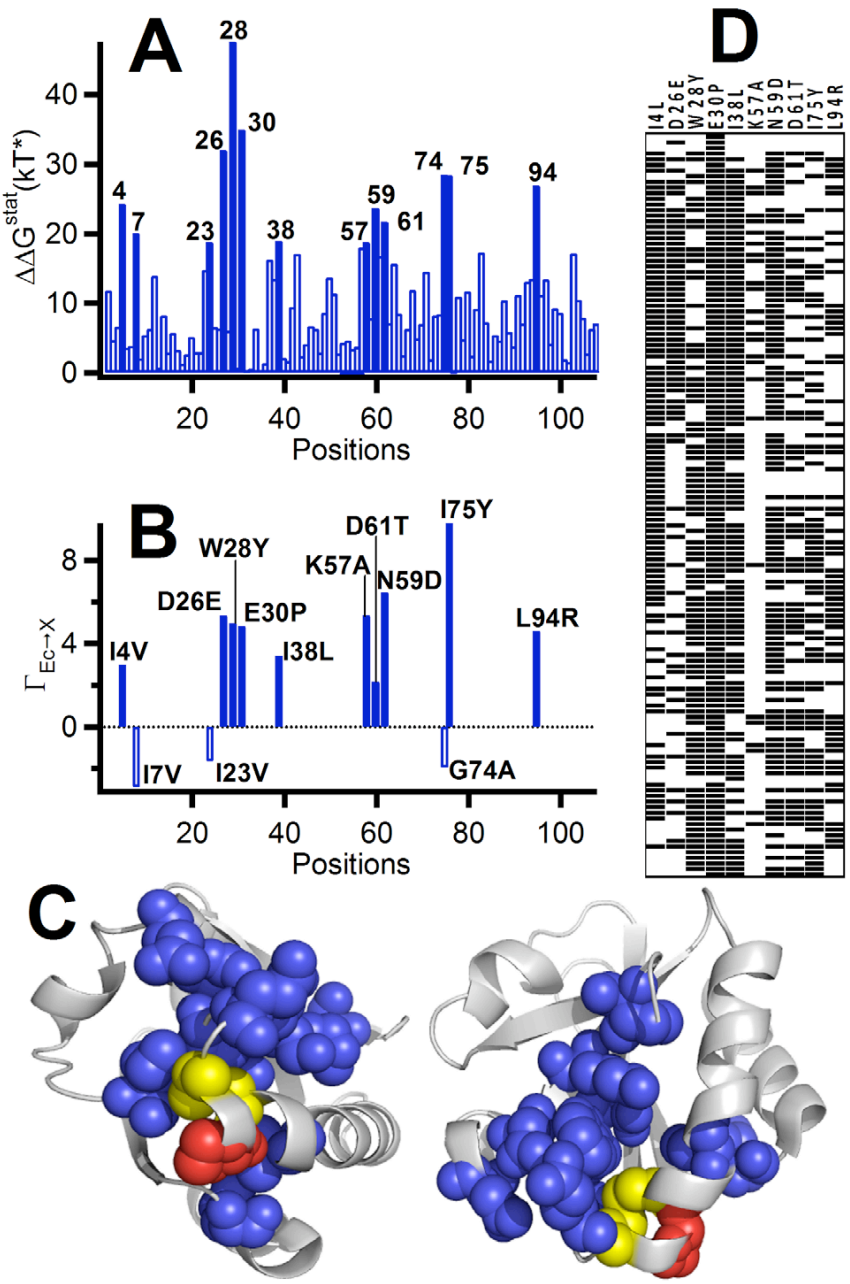

Figure 4. Statistical coupling analysis of the emergence of folding catalysis activities in thioredoxin domains. (A) Statistical free energies for the coupling of position 34 with all the other positions (*E. coli* thioredoxin numbering is used). These values have been calculated using as perturbation the signature P34H mutation. The 13 positions with the highest values for the coupling energies are labeled. (B) Determination of ten favored mutations at the positions selected in step 1 (see text for details and for the definition of the $\Gamma_{Ec\rightarrow X}$ function. (C) Mapping of the 10 positions selected for mutation (panel B) on the *E. coli* thioredoxin 3D structure (blue). The active site disulfide bridge is shown in yellow and the proline at position 34 is shown in red. (D) Occurrence of the 10 mutations selected (panel B) in the MSA sequences that include a histidine at position 34. Most of these sequences belong to eukaryotic protein disulfide isomerases. Note that many different combinations of these mutations are actually found in extant PDIs.

doi:10.1371/journal.pcbi.1002558.g004

mainly the oxidation pathway (Figure 3), an expectation that will be supported by the results reported here.

Most of the 29 variants screened show increased levels of the promiscuous activity with respect to both the background P34H variant and wt *E. coli* thioredoxin (Figure 5A). This result is consistent with the proposed role of the selected mutations in the emergence of the protein folding activities of the thioredoxin fold. What may perhaps be surprising, however, is that some of the

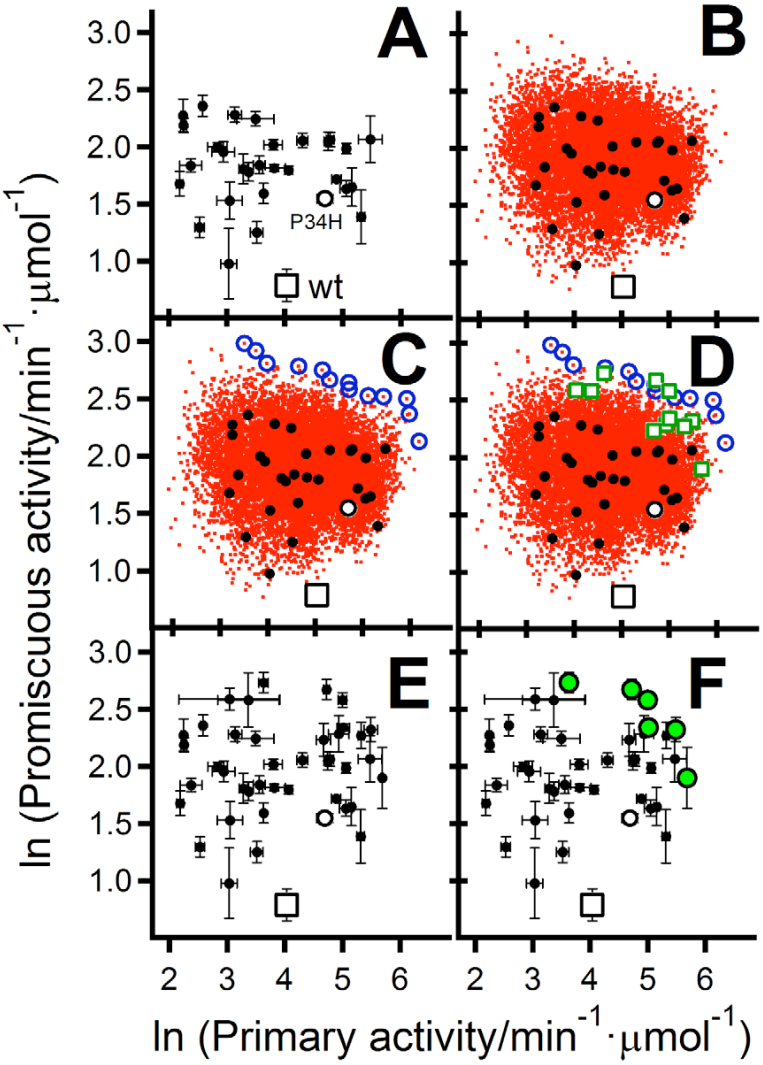

Figure 5. Approximation to the Pareto set of optimum primary/promiscuous activities. (A) Values of the primary and promiscuous activities for 29 variants randomly selected from a combinatorial library based on the 10 mutations derived from SCA analysis (see text and Figure 4). Values for the wild-type thioredoxin from *E. coli* and the background P34H variant are also included. Note that logarithms of activities are used here in and also in all the other panels of this figure. (B) Partial-least-squares (PLS) reconstruction of the data for the whole combinatorial library (small red data points). The experimental data of panel A (used as a basis for the reconstruction) are also shown, although, for the sake clarity we have omitted the errors bars here (as well as in panels C and D). The reconstructed data (red points) are actually derived from 20 bootstrapping replicas (see Methods for details). (C) Prediction of the Pareto set from the PLS-reconstructed data. The 11 variants belonging to this predicted Pareto set are shown with blue circles. (D) The actual experimental activity values for the 11 variants are shown (open green squares). (E) Expanded experimental variant set including the original 29 variants (panel A), the wild-type thioredoxin from *E. coli*, the P34H background variant and the 11 variants added as a result of PLS-reconstruction/Pareto-set-prediction (panels C and D). (F) The actual Pareto set of the expanded variant set is shown (green data points).
doi:10.1371/journal.pcbi.1002558.g005

variants also show an increased level of the primary activity, indicating the possibility of the simultaneous enhancement of the primary and promiscuous functions.

In order to assess the full range of function modulation achieved by the combinatorial library we have carried out a fit of the experimental data based on the equation:

$$\ln A^k = \sum_i \delta_i \cdot p_i^k + \sum_i \sum_{j \neq i} \delta_{ij} \cdot p_{ij}^k \qquad (2)$$

followed by a reconstruction of the entire library using the values of the fitted parameters. The meaning of the symbols in equation 2 is as follows: A^k is the dependent variable (activity) with k being a label that identifies the type of activity (i.e., k = "primary" or k = "promiscuous") = 1; δ_i is an independent variable that may take the values 0 or 1, corresponding to the absence o presence of the mutations at position i; p_i^k is a measure of the effect of the mutation at position on the activity A^k; $\delta_{ij} = \delta_i \cdot \delta_j$ is an independent variable that takes a value of 1 when mutations at positions i and j

occur simultaneously (and takes a value of zero otherwise); p_{ij}^k is a measure of the effect of the coupling between mutations at positions i and j on the activity k. Equation 2 embodies a comprehensive model that includes the effects of individual mutations (p_i^k values) as well as the possibility that mutation effects are non-additive ($p_{ij}^k \neq 0$). It involves, however, 110 fitting parameters (10 p_i^k parameters and 45 p_{ij}^k parameters for each of the two activities), while the number of experimental values to be fitted is only 58 (i.e., the values of the two activities –primary and promiscuous- for the 29 library variants studied). Having more fitting parameters than dependent variable values is a common occurrence in chemometrics, often addressed using partial least-squares [27,28] (PLS), a dimensionality reduction approach akin to principal component analysis. Indeed, the widespread usefulness of the PLS approach is often credited to its ability to handle a large number of independent variables (i.e., fitting parameters) (see chapter 7 in Livingstone [44]. PLS thus uses latent variables (latent vectors): orthogonal combinations of the original variables that explain most of the variance in the original independent variable set and are also constructed to maximize their covariance with the dependent variables. The original variables may then survive the PLS dimensionality reduction, but they are combined in a few relevant latent vectors. In the case of interest here, once a PLS fit of equation 2 to the experimental data for the 29 variants studied (Figure 5A) has been performed (see Methods for details), it is straightforward to calculate the expected primary and promiscuous activity data for the whole library of 1024 variants. Of course, there remain two important issues related to the assessment of the uncertainty associated to such full-library reconstruction and to its experimental validation. To assess reconstruction uncertainty, we have used a bootstrapping approach involving PLS fits to replica sets obtained by randomly re-sampling from the original experimental set (see Methods for details). Full-library reconstructions resulting from the PLS analyses of 20 such replicas are given in Figure 5B. They clearly suggest that the mutation set derived from the statistical coupling analysis potential has a huge potential for modulating both, the primary and promiscuous activities. Experimental validation of the reconstruction (based on experimental measurements on the predicted Pareto set of optimal variants) is described in the following section.

Approximation to the Pareto set of variants with optimal primary/promiscuous activities

We derived an "optimistic" prediction of the Pareto set of optimal primary/promiscuous activities (Figure 5C) as the set of non-dominated solutions in the ensemble of reconstructions shown in Figure 5B. The 11 variants in this predicted set were prepared and their activities determined experimentally. There is an excellent qualitative agreement between prediction an experiment, in the sense that, for all the 11 variants, increased levels of both activities were found (Figure 5D). This agreement validates the reconstruction carried out on the basis of the PLS analysis of the 29-variants set.

It is relevant to note at this point that the PLS-reconstruction/Pareto-prediction analysis leads to an expansion of our experimental variant set (from 29 variants to 40 variants) but in a manner that is not random. Actually, the 11 variants added to the experimental set allow us to move in the space of primary/promiscuous activities in the general direction of the simultaneous enhancement of both activities, as is visually apparent in Figures 5E and 5F. The Pareto set from the experimental data for the 29+11 = 40 variant set (Figures 5E and 5F) is still only an approximation to the Pareto set for the whole library, since additional cycles of PLS-reconstruction/Pareto-prediction could in principle lead to further enhancements in both activities. However, PLS-reconstruction starting with the 40-variant experimental data set suggests that additional improvements are expected to be small (see Figure 6), supporting that 40-variant Pareto set is likely close to the Pareto set of the full library. Furthermore, the main result of the analysis is already apparent with the 40-variant set: *E. coli* thioredoxin, as well as the background P34H variant for library construction, does not belong to the Pareto set and, therefore, simultaneous enhancement of the primary and promiscuous activities is feasible (and has been experimentally achieved: Figures 5E and 5F). Note that, in addition to the targeted simultaneous enhancement (implying enhanced levels for both activities), the experimental data set (as well as the PLS reconstructions of the full combinatorial library) indicates that the two activities can be modulated in an independent-like manner and includes "specialist" variants with a high level for one activity and low value for the other.

How large are the primary/promiscuous activity modulations achieved?

Figure 7A highlights the modulation ranges experimentally achieved for the reductase and catalysis of oxidative folding activities (about 33-fold and 7-fold, respectively). One important issue is whether these ranges (in particular that of the promiscuous activity) are to be considered large or small. The answer to this question depends largely on the relevant context. Certainly, the modulation ranges we have found are much smaller than those

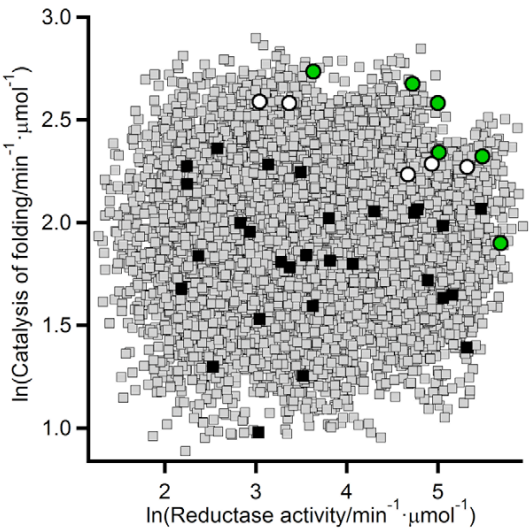

Figure 6. Full-library partial-least-squares reconstructions of the reductase/catalyis-of-oxidative-folding data based upon the expanded 40-variant set. The black squares represent the first-round 29-variant set and the circles represent the 11 variants added as a result of the experimental validation of the Pareto prediction shown in Figure 5C (green data points are used here for the Pareto set). Error bars have been omitted for clarity, but they are shown in Figures 5E and 5F. The reconstructed data (grey squares) are derived from 20 bootstrapping replicas of the 40-variant data. Note that the PLS reconstructions shown here are based on the expanded 40-variant set, while those of Figure 5B were based on the first-round 29-variant set. The full-library reconstructions shown here suggest only small enhancements over the 40-variant Pareto set would be obtained in additional screening rounds and, further, they support that the 40-variant Pareto set is already close to the full-library Pareto set.
doi:10.1371/journal.pcbi.1002558.g006

reported in some protein design studies (consider, for instance, the 200-fold enhancement in engineered Kemp eliminase activity reported by Baker, Tawfik and coworkers [45]) and the ranges typically considered relevant in a biotechnological application context. It is important to note, however, that the approach we have used is aimed at assessing the patterns of primary/promiscuous activity modulation in the mutational space actually explored by natural evolution when recruiting the promiscuous activity for a new function. That is, the mutations included in our combinatorial library are those expected to be associated to the emergence of folding catalysis in the thioredoxin scaffold during the course of natural evolution. If the approach is successful, promiscuous activities approaching the levels of natural thioredoxin-scaffold folding catalysts should be reached. In an evolutionary/biological context, therefore, the promiscuous activity modulation achieved should be compared with the evolutionary significant modulation range estimated on the basis of activity data for a protein disulfide isomerase. Experimental data for bovine PDI are included in Figure 7 and indeed show an acceptable level of congruence with the Pareto set for the 40-variant variant set (Figure 7A) as well as with the corresponding PLS reconstructions (Figure 7B).

Clearly, the modulation range achieved for the catalysis of oxidative folding is significant from a biological/evolutionary point of view. Interestingly, this does not appear to hold for the other folding-related activity of thioredoxin domains: the disulfide-reshuffling activity responsible for the rescue of misfolded proteins with incorrect disulfide bridges. Figure 7C is a plot of reshuffling activity (measured using disulfide-scrambled ribonuclease A as substrate: see Methods for details) versus catalysis of oxidative folding, including the 40-variant set, wild-type thioredoxin from *E. coli*, the P34H background variant and bovine PDI. Figure 7D is a similar plot including the PLS reconstructions based on the experimental data of Figure 7C. These two plots suggest that the combinatorial library used (based on the mutation set derived from SCA analysis: Figure 4) spans the evolutionary relevant range for the catalysis of oxidative folding, but not the corresponding range for the reshuffling activity. This result is actually consistent with some known features of the structure-function relationship in protein disulfide isomerases. PDIs have a multidomain structure usually described [46] in terms of four distinct domains (a b b′ a′), two of which (the a and a′ domains) display the thioredoxin-fold structure and the CXXC active site motif responsible for the catalysis of disulfide-linked process. The isolated a and a′ domains have been shown to introduce efficiently disulfide bridges into proteins [47], while additional domains are required for efficient catalysis of disulfide bond reshuffling in folded proteins [48–50], perhaps because the "inactive" b and b′ domains play a role in facilitating steps that involve difficult conformational changes [48]. Obviously, this "multidomain" effect cannot be reproduced by engineering based on a single-domain thrioredoxin scaffold.

The shape of the Pareto set in the primary/promiscuous activity plot is consistent with the conformational diversity hypothesis

The primary/promiscuous plots presented so far (Figures 5, 6 and 7), employ logarithmic activity scales in order to emphasize the order of magnitude of the modulations achieved. However, using linear scales in these plots (Figures 8A and 8B) reveals a surprisingly simple pattern: a linear-like Pareto set and a roughly triangular shape for the "cloud" of experimental data points below the Pareto set. This pattern is robust, being observed in the 40-variants data set and in the PLS reconstructions of the full combinatorial library. Note also that the observed pattern implies that essentially all the experimental data points are at or below a line connecting the expected maximum values for the primary and promiscuous activities and, therefore, that the experimental data points populate an area in the primary/promiscuous activity plot which is about half the maximum area accessible. The probability of this happening by chance if there is no correlation between the primary and promiscuous activities is on the order of (1/2)^NDP being NDP the number of experimental data points. This gives a negligible probability for NDP = 1024 even for NDP = 40. Finally, as we have already pointed out, the second-round of the library screening process was sharply focused to the Pareto set and that, as a result, the Pareto set of the 40-varaints experimental set is likely to be close to the Pareto set of the full library. We conclude from all this reasoning that the simple experimental pattern in Figures 8A and 8B is robust and is unlikely to have arisen by chance. It is natural then to seek a simple interpretation for such a simple, but intriguing pattern. As we elaborate below one simple explanation is provided by the so-called conformational diversity hypothesis.

The conformational diversity hypothesis posits native proteins may exist in solution as different conformations in equilibrium and provides a plausible structural rationale for the existence of protein promiscuity [5,51,52]. In very simple terms, the most populated (i.e., dominant) conformation is responsible for the primary activity while alternative, low-population conformations perform the promiscuous activities. Mutations can shift the equilibria between the different conformations and thus modulate the balance between the primary and promiscuous activities. A linear-like Pareto set could thus be explained in terms of two optimal conformations, each being responsible for catalyzing efficiently only one of the activities. For instance, one conformation would achieve molecular optimization for the substrate reduction process (Figure 2) when the active-site disulfide is reduced, while an alternative conformation would achieve optimization for substrate oxidation (Figure 3) when the active-site disulfide is oxidized. Obviously, data points below the Pareto set would correspond to significant population of other conformations that are suboptimal in terms of activity. This interpretation is clarified below with a simple illustrative example.

Consider three protein conformations: $\mathbf{a_0}$: a conformation with no activity; $\mathbf{a_1}$: the conformation responsible for the primary activity; $\mathbf{a_2}$: the conformation responsible for the promiscuous activity. The mol fractions of the three conformations must add up to unity:

$$X(\mathbf{a}_0)+X(\mathbf{a}_1)+X(\mathbf{a}_2)=1 \quad (3)$$

Mutations may change these mol fractions and, obviously, the optimal primary/promiscuous activity situations will be achieved when $X(\mathbf{a_0})=0$ and,

$$X(\mathbf{a}_1)+X(\mathbf{a}_2)=1 \quad (4)$$

Since activities should be proportional to the corresponding mol fractions, equation 4 defines a straight line in a plot of promiscuous versus primary activity. We refer to this line as the "trade-off line". In the same plot, suboptimal situations in which $X(\mathbf{a_0})\neq 0$ will necessarily be represented by points in a triangular area defined by the trade-off line and the plot axes (see Figure 8D). Certainly, the plot in Figure 8D (showing the optimal trade-off line and a shaded triangular region corresponding to suboptimal situations) is an idealized representation. In practice, we must consider the possibility that the mutations used are unable to completely shift

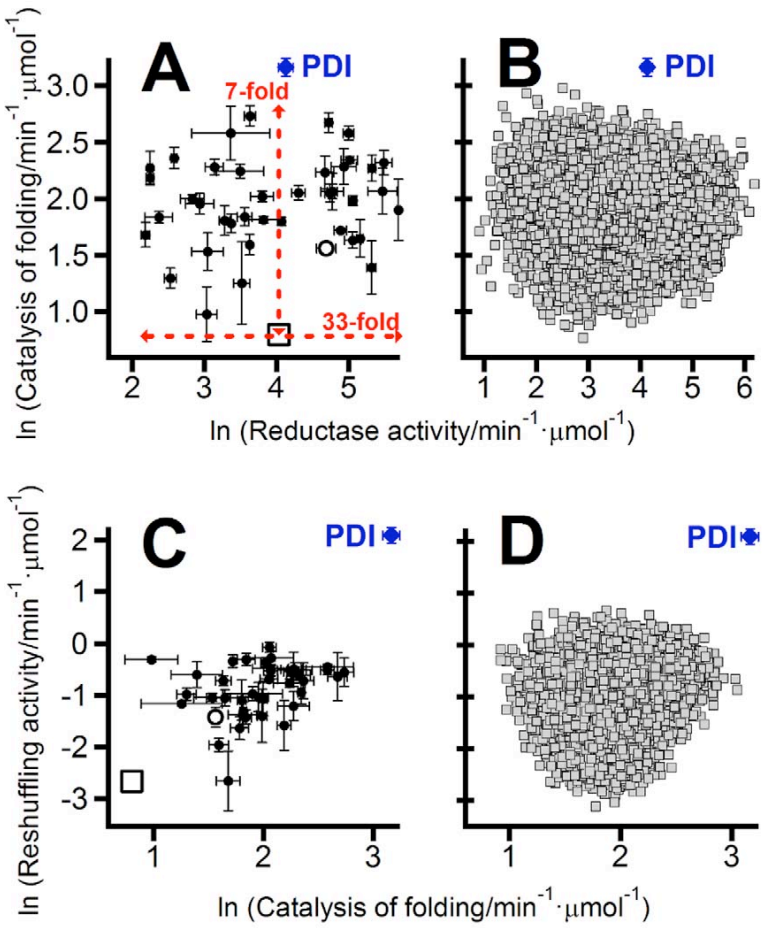

Figure 7. Assessment of the modulation ranges achieved for the primary and promiscuous activities. (A) Experimental reductase and catalysis of oxidative folding data for the expanded 40-variant set, the wild-type thioredoxin from *E. coli* and the background P34H variant. The meaning of symbols is as in Figure 5A. The approximate modulation ranges achieved are indicated with arrowed red lines. Data of bovine protein disulfide isomerase (blue data point labeled PDI) are also included to provide an evolutionary relevant scale for comparison. Note that specific activity data for PDI are given per molar concentration of active thioredoxin domain. (B) Comparison of the PDI experimental data with the partial least squares reconstruction of the reductase/catalysis-of-oxidative-folding data for the whole combinatorial library. The reconstructed data (grey squares) are actually derived from 20 bootstrapping replicas. (C) Experimental disulfide-reshuffling activity and catalysis of oxidative folding data for the expanded variant set, the wild-type thioredoxin from *E. coli* and the background P34H variant. Data of bovine PDI are also included to provide an evolutionary relevant scale for comparison. (D) Comparison of the PDI experimental data with the partial least squares reconstruction of the reshuffling/catalysis-of-oxidative-folding data for the whole combinatorial library. The reconstructed data (grey squares) are actually derived from 20 bootstrapping replicas.
doi:10.1371/journal.pcbi.1002558.g007

the equilibria towards the active conformations (i.e., they might be unable make the mole fraction of the inactive conformation strictly equal to zero). To provide an illustration of this situation, we have carried out a stochastic simulation of a 40-variant data set, assuming that conformation populations are proportional to statistical weights derived from flat distributions. That is, the mol fraction of a given conformation a_i is given by $w_i/\Sigma w_i$ where w_i is its statistical weight (derived from a random number generator in the [0,1] interval) and Σw_i is the sum of the statistical weights for all the conformations. The result of this simulation (Figures 8E) is a roughly triangular-shaped cloud of data points with a linear-like Pareto set that approaches the trade-off line.

For simplicity and illustration, the simulations included in Figure 8 assume that there is only one sub-optimal conformation and that it has zero primary and promiscuous activity levels. It is important to note, however, that the general result of the simulations is robust and it is obtained with different conformation models (see Figures 9A–C) including several sub-optimal conformations with non-zero activity levels. Apparently, all that is required for a linear-like Pareto set to be obtained in these simulations is a model with two optimal conformations with quite different capabilities to catalyze the primary and promiscuous processes. Actually, if the two optimal conformations are efficient at catalyzing both activities (and, therefore, there are no trade-offs!), the pattern of a linear-like Pareto set with a triangular-like data points cloud is not obtained (see Figure 9D for a representative simulation).

Certainly, the simulations discussed above (Figures 8 and 9) are not meant to be taken as direct evidence in support of the conformational diversity. In fact, obtaining such direct evidence

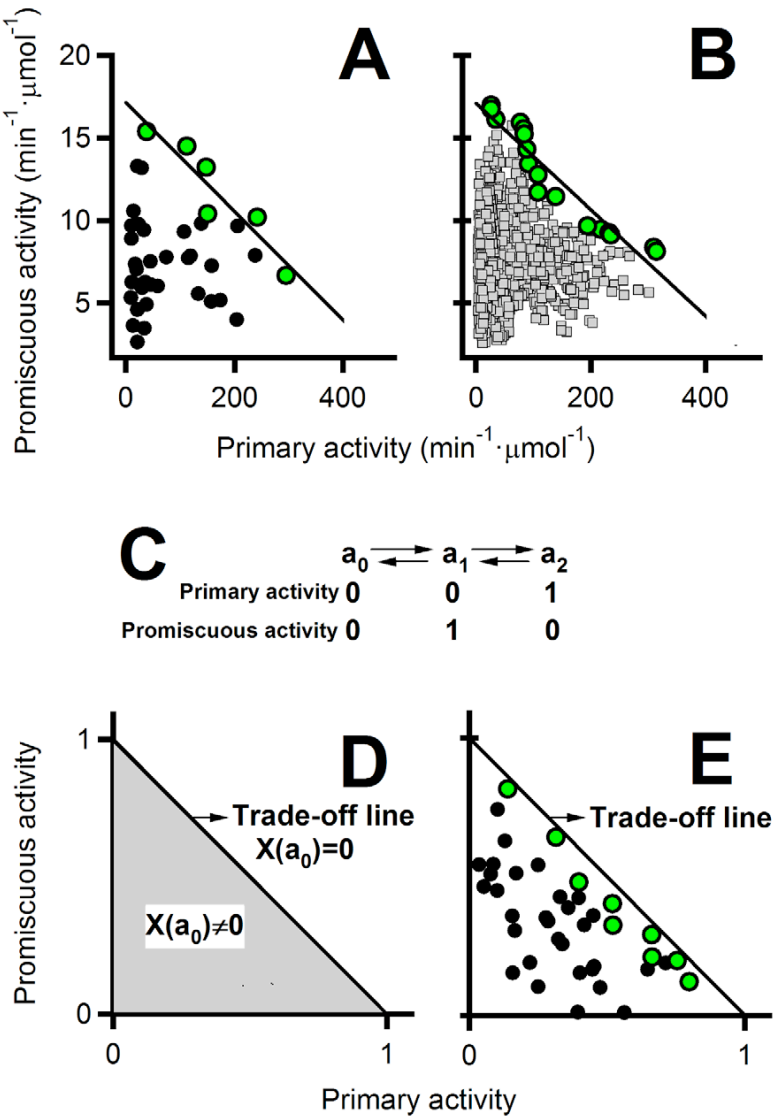

Figure 8. The shape of the primary/promiscuous Pareto set is consistent with the conformational diversity hypothesis. (A) Values of the reductase and catalysis of oxidative folding activities for the 40-variant experimental set. The Pareto set is shown with green data points. This plot is analogous to that in Figure 5F but linear activity scales (instead of logarithms) are used here and error bars have been omitted for the sake of clarity. The line represents the least-squares fit to the Pareto set data. (B) A partial least squares reconstruction of the data for the whole combinatorial library based on the experimental data of panel A. The data shown correspond to a single bootstrapping replica; however, other replicas show the same general pattern The Pareto set (green points) and the corresponding linear fit are shown. (C) Simple conformational diversity model used in the simulation summarized in panels D and E. Conformation a_0 is not active, while conformations a_1 and a_2 are responsible for the promiscuous and primary activities, respectively. (D) Relevant regions in the primary/promiscuous activity diagram according to the model shown in panel C. Optimal situations correspond to a zero mol fraction of the inactive conformation and define a *trade-off line* in the plot. Sub-optimal situations correspond to a mol fraction of a_0 higher than zero and are located in a triangular-shaped region below the trade-off line. (E) Stochastic simulation of 40-variant set based on the model shown in panel C. The Pareto set (green data points) approaches the trade-off line.
doi:10.1371/journal.pcbi.1002558.g008

would require extensive structural and dynamic characterization (see, for instance [53] and [54]) which is beyond the scope of this work. Because of this we cannot rule out that other kinds of models (based, for instance, on modeling the mutation effects on the activities) could also explain the experimental results founds. Nevertheless, it is clear that the conformational diversity hypothesis provides a simple, Occam-razor explanation (since modeling of specific mutation effects is not involved) for an equally simple, but otherwise intriguing, experimental modulation pattern.

Diversity of patterns of primary/promiscuous activity modulation

The simultaneous enhancement of the primary and promiscuous activities we have discussed in the preceding sections is only

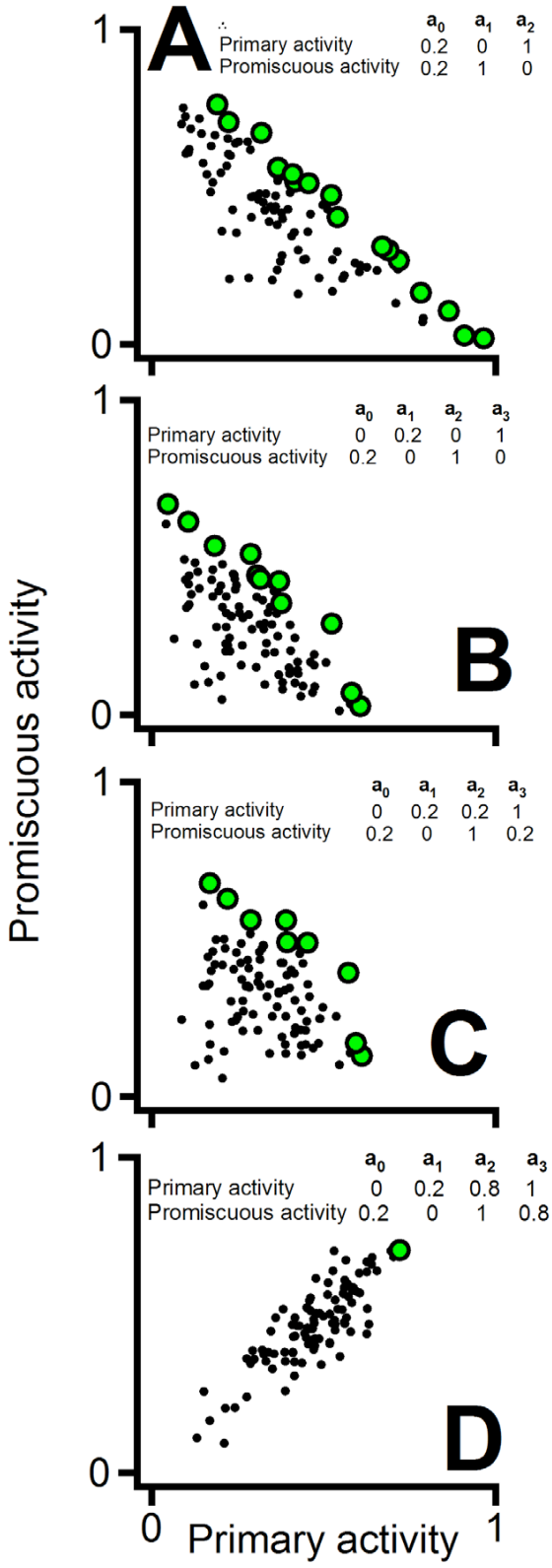

Figure 9. Stochastic simulations of primary/promiscuous activities based on different conformational diversity models. (A) Simulation based on a 3-conformation model identical to that used in the simulation of Figure 8E, except that a low (but different from zero) level of primary and promiscuous activities for the suboptimal conformation ($\mathbf{a_0}$) is assumed. Here, as well as in the other panels, the Pareto set is shown with green circles. (B) Simulation based on a 4-conformation model with two suboptimal conformations. (C) Same as in (B), except that each one of the optimal conformations ($\mathbf{a_2}$ and $\mathbf{a_3}$) has a low level of the alternative activity. (D) Same as in (C), but assuming that the optimal conformations show high level of both activities. The models used in (A), (B) and (C) include the existence of trade-offs between the optimal conformations and yield roughly linear Pareto sets with a significant number of data points below the Pareto set due to the presence of suboptimal conformations. This is actually the experimental pattern we have found for the catalysis-of-oxidative-folding/reductase activities (Figure 8A). The model used in (D) does *not* include trade-offs between the optimal conformations (i.e., $\mathbf{a_2}$ and $\mathbf{a_3}$ efficiently catalyze both, the primary and the promiscuous processes) and the pattern is completely different: a significant correlation between the two activities and a very small Pareto set are observed.
doi:10.1371/journal.pcbi.1002558.g009

one aspect (albeit a prominent one) of a general property of the set of mutations derived from the function-based statistical coupling analysis: the potential for originating a multiplicity of mutational paths leading to different types of function modulation patterns. To illustrate the idea (Figure 10) we use one of the full-library reconstructions derived from the PLS analysis of the 40-variants experimental set. Each of the mutational paths shown in Figure 10 has been constructed using the following procedure: a) An initial variant is chosen; b) the variants connected to the chosen one by single mutations are tested for a given activity-related condition; c) one variant among those that pass the test is randomly selected; d) the cycle a-c is repeated until no mutational steps are available.

The paths shown in Figure 10A start with the background variant (i.e., the variant with no mutations) and mutational steps are allowed if promiscuous activity (catalysis of oxidative folding) is increased while the primary activity (reductase) is maintained above a certain threshold. These simulations illustrate the case in which there is selection for enhanced promiscuous activity while maintaining a level of the primary activity that does not compromise fitness. Several paths lead to a variant with increased promiscuous activity and still a significant level of primary activity. Interestingly, PDIs show a significant level of reductase activity (see published work [47,48,50] and Figure 7), perhaps because the catalysis of disulfide-linked folding likely involves steps in which incorrect disulfide bridges must be broken up. It is thus tempting to speculate that the no-trade-off paths in Figure 10A illustrate some of the actual function changes taking place in the evolution of these disulfide-linked folding catalysts. It is also interesting that some of the intermediate variants in the paths of Figure 10A have significantly increased levels of both activities, again emphasizing the possibility of the simultaneous enhancement of the primary and promiscuous functions.

The Paths shown in Figure 10B use as starting point a variant with comparatively high values of both, the primary activity and the promiscuous activity (actually, a member of the Pareto set of optimal solutions) and mutational steps are allowed if the primary activity is decreased while the promiscuous activity remains above a given threshold. The paths in Figure 10B lead to a specialist protein with high promiscuous activity and low primary activity. These paths could be viewed as illustrating the molecular changes that could in some cases occur in one of the gene copies arising from the gene duplication event involved in the emergence of a new function. According the so-called balance hypothesis [55], single-gene duplication may actually be harmful because it

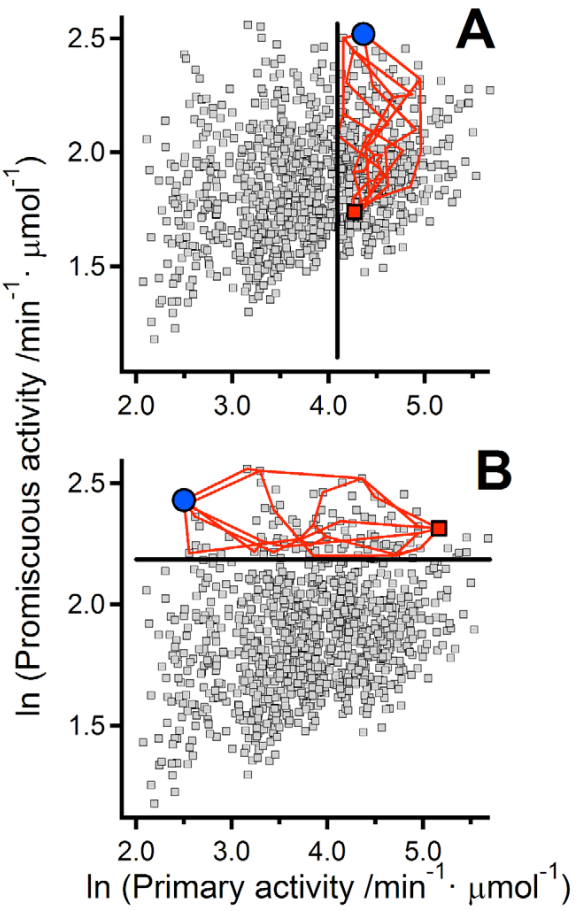

Figure 10. Diversity of mutational paths in the primary/promiscuous activity space. Grey data points represent the results of a PLS reconstruction based on the 40-variant experimental set. Mutational paths (red lines) connecting a starting variant (red data point) and a final variant (blue data point) are shown. (A) The starting variant shows a low level of the promiscuous activity and a comparatively high level of the primary activity. Mutational steps are allowed if promiscuous activity is increased and the primary activity is maintained above a given threshold (shown as a vertical black line). (B) The starting variant shows high level of both, the primary and promiscuous activities. Mutational steps are allowed if the primary activity is decreased and the promiscuous activity is maintained above a given threshold (shown as a horizontal black line).
doi:10.1371/journal.pcbi.1002558.g010

immediately leads to a very large excess of a given protein, which may be deleterious. If imbalance is associated to an excessive level of the primary function, then the mutational paths in Figure 10B illustrate a potential of the mutational space explored by natural selection to efficiently restore balance.

A final clarification should be made. The mutational paths in the illustrative simulations summarized in Figure 10 have been obtained assuming that all the single-mutation steps can be readily achieved, although some of them cannot be realized with a single-base substitution. However, these amino acid substitutions do occur during natural evolution (involving an intermediate amino acid) as clearly shown by the sequences in Figure 4D. In connection with this, it is important to note that the mutational space we have characterized is very likely a subspace of the full mutational space explored by natural selection in the evolution of disulfide-linked folding catalysts (the latter involving additional positions and several mutations at each position). Obviously, this fact only reinforces our conclusions.

Concluding remarks

Current views on the relation between primary and promiscuous protein activities are derived to a significant extent from laboratory evolution experiments aimed at enhancing promiscuous functions. Many of these studies have found a decrease (often moderate) in primary activity concomitantly with the increase in promiscuous function, suggesting that the two activities trade-off. In this work, we have introduced an approach to determine how the interplay between the primary and promiscuous activities of a protein is modulated in the mutational space evolutionary linked to the emergence of a new function. Application of this new approach to the emergence of folding catalysts reveals a hitherto unexplored scenario: diverse patterns of primary/promiscuous activity modulation may occur as response to different types of evolutionary pressure, including no-trade-off paths involving the simultaneous enhancement of both activities. Some general remarks related with this result are appropriate:

1) Although admittedly moderate, the simultaneous enhancement we have achieved is in the range expected for modulations that are significant in a biological context, as indicated by the acceptable level of congruence of PDI data with the Pareto set for the 40-variant variant set (Figure 7A) as well as with the corresponding PLS reconstructions (Figure 7B).
2) We have used in this work combinatorial-library screening to probe the mutational interplay between primary and promiscuous activities as assessed by established *in vitro* assays, while the interplay between the two activities *in vivo* may also be determined by additional (non-mutational) factors, such as the redox environment [56]. It must be noted, however, that the mutations included in our library were derived from a coupled correlation analysis (SCA) targeted to the evolutionary emergence of folding catalysis in the thioredoxin scaffold and, further, that combinations of these mutations do enhance this activity in *E. coli* thioredoxin to levels approaching those of the natural folding catalysts (PDI data: see Figure 7A). There can be little doubt, therefore, that the mutational modulations captured by our approach have an evolutionary significance related with new function emergence. In particular, the possibility arises that processes of new function evolution could actually involve no-trade-off mutational paths prior to the gene duplication event.
3) Contrary to naïve intuition, the fact that variants with simultaneously enhanced levels of primary an promiscuous activities can be obtained is not necessarily in conflict with the existence of trade-offs between these two activities, because trade-offs are confined to the Pareto set of optimal primary/promiscuous solutions while most of the variants spanned by the mutational space studied do not belong to this Pareto set.
4) Obviously, the simultaneous enhancement of primary and promiscuous activities relies on the fact that the maximum possible values of the two activities are not realized in the variant used as background of the mutational analysis. The implication is that the background, wild-type protein is not only sub-optimal for its promiscuous activity (as was to be expected), but that it is also sub-optimal for its primary activity. Interestingly, a very recent analysis of the catalytic parameters for several thousand enzymes [57] indicates that

most enzymes have moderate catalytic efficiency and supports that primary-activity enhancement should be possible in many cases. The scenario we have described in this work for the mutational interplay between primary and promiscuous activities could then be expected to be widespread, and not limited to the specific type of protein promiscuity (forward/reverse reactions) we have experimentally studied.

5) The partial-least-squares/Pareto-set-prediction protocol we have used provides an efficient computational basis to library screening targeted at the in vitro enhancement of several protein activities. The general result obtained using this protocol will of course depend on the type of library being screened. Libraries focused on the basis of sequence alignment information (such as that used in this work) will likely lead to the comparatively small enhancements that are relevant in a biological context and that may be informative about the evolution of protein function. However, libraries focused on the basis of successful computational design and very large random libraries may lead to the large enhancements that are significant in a biotechnological setting. Therefore, the availability of partial-least-squares/Pareto-set-prediction protocol, together with the general description introduced here for the interrelation between the several activities of a protein and the mutational space explored, should pave the way for the engineering of multi-functional enzymes.

Methods

Sequence alignment and Statistical Coupling Analysis

BLAST2 (1996–2003, W. Gish http://blast.wustl.edu) was used to search the TrEMBL sequence database of October-2007 (http://www.ebi.ac.uk/trembl) using the sequence of *E. coli* thioredoxin as query. The resulting sequences were aligned with the query sequence using the Smith-Waterman algorithm and only those with sequence identity with the query of 0.3 or higher were retained for further analysis. We made no further attempt to correct or filter the alignment, since the results obtained from its analysis made clear sense from both, the structural and functional viewpoints (see Figures 4 and 5). Of the1440 sequences in the alignment used, 1264 had a proline at position 34 and 132 had a histidine at that position. Essentially all the sequences with histidine at position 34 belonged to eukaryotes and most of them were actually annotated as protein disulfide isomerases.

Statistical coupling analysis of the sequence alignments based on the P34→H perturbation were performed using homemade programs, but in a manner identical to that described by Lockless and Ranganathan [24]. The robustness of this analysis is supported by the fact that the positions with high values for the statistical coupling energy also rank high in a simple covariance analysis [58] of the sequence alignments (see Figure S1 in Supporting Information).

Variant library generation and protein purification

The combinatorial library of thioredoxin variant sequences on the P34H background was prepared by using gene assembly mutagenesis as we have previously described [59]. For ease of protein purification, the genes encoded a His_6 tag at the N-terminal end (i.e., at a position roughly opposite to the active-site region). Purification of the thiredoxin variants, assessment of their purity and concentration measurements, were performed as previously described [59]. Bovine PDI was purchased from Sigma and used without further purification.

Activity determinations

Reductase activity of the thioredoxin variants was determined at 37°C by a turbidimetric assay of the thioredoxin catalyzed reduction of insulin [60]. Briefly, thioredoxin-variant (or PDI) solutions at pH 6.5 (phosphate buffer 0.1 M) in the presence of 2 mM EDTA and 0.5 mg/mL insulin were prepared. The reactions were initiated by addition of DTT to a 1 mM final concentration and monitored by measuring the absorbance at 650 nm (A_{650}) as function of time. Activity is calculated as the maximum value of the change of A_{650} with time, i.e., the maximum value for the derivative dA_{650}/dt (see Figures S2 and S3 in Supporting Information for representative examples). Typically, for each variant, 3–4 experiments at different thioredoxin-variant concentrations (within the 0–5 μM range) were carried out and the specific activity values, together with its associated standard errors, were determined from linear fits to the activity versus concentration profiles: see Figure S4 in Supporting Information for representative examples and for further details.

The catalysis of oxidative folding activity was determined by following the recovery of ribonuclease A (RNase A) activity from completely reduced RNase A following the procedure described by Lundström et al. [43]. Briefly, nitrogen-saturated solutions of thioredoxin variants (or PDI) in 0.1 M phosphate buffer pH 7 in the presence of 1 mM EDTA and 100 μM oxidized glutathione were prepared. The reaction was initiated by addition of RNase A from a stock solution to a final concentration of 0.4 mg/mL. After 1 hour incubation at 37°C, the RNase A activity was determined using the standard assay based on the hydrolysis of 2′-3′-cCMP. Typically, for each variant, 4 experiments at different thioredoxin-variant concentrations (within the 0–15 μM range) were performed and the specific activity values, together with its associated standard errors, were determined from linear fits to the recovered RNase A activity versus concentration profiles. Assays for the disulfide reshuffling activity were carried out in the same way, except that disulfide-scrambled RNase A was used and 100 μM reduced glutathione was included in the reaction solution. Fully-reduced and scrambled RNase A were prepared as described by Lündstrom et al. [43] See Figures S5 and S6 in Supporting Information for representative examples of the disulfide-linked folding assays and for further details.

Partial least-squares analysis

PLS analyses were carried out with the program Unscrambler X from CAMO software using the NIPALS algorithm. In all cases, the dependent variables were the logarithms of the values for the primary and promiscuous activities and were auto-scaled (i.e., they were subjected to mean subtraction followed by division by the standard deviation) prior to the analysis. Leave-one-out cross-validation was used and the number of latent variables retained was the optimum value suggested by the Unscrambler program on the basis of the mean square error of cross-validation. Actually, the PLS analyses were carried with 20 replica sets constructed from the original set through random re-sampling (bootstrapping) and the number of latent variables retained did depend on the replica set used; typical values, however, were on the order of 3–11 (i.e., much smaller than the numbers of dependent and independent variables involved). Illustrative plots experimental versus predicted activities are given in Figure S7 in Supporting Information.

Supporting Information

Figure S1 Comparison between the statistical free energies derived from SCA analysis and the results of a simple covariance analysis (σ values). The values shown correspond to the correlation of position 34 with all other positions

in the thioredoxin sequence. The 13 positions with the highest statistical free energy values (see Figure 4A in the main text) are shown here with closed circles.
(TIF)

Figure S2 Determination of the reductase activity of thioredoxin variants. Respresentative plots of absorbance at 650 nm versus time for the reduction of insulin catalyzed by thioredoxins. Profiles for the wild-type thioredoxin from *E.coli* and two variants are shown.
(TIF)

Figure S3 Determination of the reductase activity of thioredoxin variants. Plots of derivative of absorbance versus time corresponding to the profiles shown in Figure S2. The activity value for each variant at the concentration shown is calculated as the maximum value of dA_{650}/dt.
(TIF)

Figure S4 Determination of the reductase activity of thioredoxin variants. Specific activity is determined from the slopes of plots of activity (maximum value of dA_{650}/dt) versus protein concentration. Several representative examples are shown, including wild-type thioredoxin from *E.coli* and several variants.
(TIF)

Figure S5 Determination of the catalysis of oxidative folding activity of thioredoxin variants. Plots of recovered RNase activity from fully reduced protein versus thioredoxin variant concentration. Specific activity is calculated as the slope of these plots. Recovered RNase activity is measured by the initial rate of the change of the absorbance at 288 nm that accompanies the hydrolysis of 2′-3′-cCMP. Several representative examples are shown, including wild-type thioredoxin from *E.coli* and several variants, one of which is the P34H variant used as background for library construction.
(TIF)

Figure S6 Determination of the disulfide reshuffling activity of thioredoxin variants. Plots of recovered RNase activity from disulfide-scrambled protein versus thioredoxin variant concentration. Specific activity is calculated as the slope of these plots. Recovered RNase activity is measured by the initial rate of the change of the absorbance at 288 nm that accompanies the hydrolysis of 2′-3′-cCMP. Several representative examples are shown, including wild-type thioredoxin from *E.coli* and several variants, one of which is the P34H variant used as background for library construction.
(TIF)

Figure S7 Representative examples of partial least squares fits to the 29-variants set. (Figure 5A). Actually, fits to 4 bootstrapping replicas (color-coded) extracted from that set are shown.
(TIF)

Author Contributions

Conceived and designed the experiments: JMSR. Performed the experiments: HGS. Analyzed the data: HGS. Contributed reagents/materials/analysis tools: BIM. Wrote the paper: JMSR. BIM supervised HGS.

References

1. Jensen RA (1976) Enzyme recruitment in evolution of new function. Annu Rev Microbiol 30: 409–425.
2. Copley SD (2003) Enzymes with extra talents: moonlighting functions and catalytic promiscuity. Curr Opin Chem Biol 7: 265–272.
3. Khersonsky O, Roodveldt C, Tawfik DS (2006) Enzyme promiscuity: evolutionary and mechanistic aspects. Curr Opin Chem Biol 10: 498–508.
4. Babtie A, Tokuriki N, Hollfelder F (2010) What makes an enzyme promiscuous?. Curr Opin Chem Biol 14: 200–207.
5. Khersonsky O, Tawfik DS (2010) Enzyme promiscuity: a mechanistic and evolutionary perspective. Annu Rev Biochem 79: 471–505.
6. Bornscheuer UT, Kazlauskas RJ (2004) Catalytic promiscuity in biocatalysis: using old enzymes to form new bonds and follow new pathways. Angew Chem Int Ed 43: 6032–6040.
7. Kazlauskas RJ (2005) Enhancing catalytic promiscuity for biocatalysis. Curr Opin Chem Biol 9: 195–201.
8. Yoshikuni Y, Ferrin T, Keasling JD (2006) Designed divergent evolution of enzyme function. Nature 440: 1078–1082.
9. Bloom JD, Romero PA, Lu ZL, Arnold FH (2007) Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. Biol Direct 2: 17.
10. Bloom JD, Arnold FH (2009) In the light of directed evolution: pathways of adaptive protein evolution. Proc Natl Acad Sci U S A 106: 9995–10000.
11. Nobeli I, Favia AD, Thornton JM, (2009) Protein promiscuity and its implications for biotechnology. Nat Biotechnol 27: 157–167.
12. Gerlt JA, Babbitt PC (2009) Enzyme (re)design: lessons from natural evolution and computation. Curr Opin Chem Biol 13: 10–18.
13. D'Ari R, Casadesús J (1998) Underground metabolism. Bioessays 20: 181–186.
14. Kim J, Copley SD (2007) Why metabolic enzymes are essential or nonessential for growth of *Escherichia coli* K12 on glucose. Biochemistry 46: 12501–12511.
15. Erijman A, Aizner Y, Shifman JM (2011) Multispedific recognition: mechanism, evolution, and design. Biochemistry 50: 602–611.
16. Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. Proc Biol Sci 256: 119–124.
17. Aharoni A, Gaidukov L, Khersonsky O, Gould SM, Roodveldt C, et al. (2005) The 'evolvability' of promiscuous protein functions. Nat Genet 37: 73–76.
18. O'Brien PJ (2006) Catalytic promiscuity and the divergent evolution of DNA repair enzymes. Chem Rev 106: 720–752.
19. Conant GC, Wolfe KH (2008) Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet 9: 938–950.
20. McLoughhlin SY, Copley SD (2008) A compromise required by gene sharing enables survival: implications for evolution of new enzyme activities. Proc Natl Acad Sci U S A 105: 13497–13502.
21. Du J, Say RF, Lü W, Fuchs G, Einsle O (2011) Active-sit remodeling in the bifunctional fructose-1,6-bisphosphate aldolase/phosphatase. Nature 478: 534–537.
22. Fushinobu S, Nishimasu H, Hatorri D, Song H.-J, Wakagi T (2011) Structural basis for the bifunctionality of fructose-1,6-bisphosphate aldolase/phosphatase. Nature 478: 538–541.
23. Babtie AC, Bandyopadhyay S, Olguin LF, Hollfelder F (2009) Efficient catalytic promiscuity for chemically distinct reactions. Angew Chem Int Ed 48: 3692–3694.
24. Lockless SW, Ranganathan R (1999) Evolutionary conserved pathways of energetic connectivity in protein families. Science 286: 295–299.
25. Lee J, Natarajan M, Nashine VC, Socolich M, Vo T, et al. (2008) Surface sites for engineering allosteric control in proteins. Science 322: 438–442.
26. Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, et al. (2005) Evolutionary information for specifying a protein fold. Nature 437: 512–518.
27. Abdi H (2003) Partial least squares regression. In: Lewis-Beck M, Bryman A, Futing M, eds. The SAGE Encyclopedia of Social Sciences Research Methods. pp. 1–7.
28. Fox RJ, et al. (2007) Improving catalytic function by ProSAR-driven enzyme evolution. Nat Biotechnol 25: 338–344.
29. Suarez M, Tortosa P, Carrera J, Jaramillo A (2008) Pareto optimization in computational protein design with multiple objectives. J Comput Chem 29: 2704–2711.
30. Suarez M, Tortosa P, Garcia-Mira MM, Rodriguez-Larrea D, Godoy-Ruiz R, et al. (2010) Using multi-objective computational design to extend protein promiscuity. Biophys Chem 147: 13–19.
31. He L, Friedman AM, Balley-Kellog C (2012) A divide-and-conquer approach to determine the Pareto frontier for optimization of protein engineering experiments. Proteins 80: 790–806.
32. Mamathambika BS, Bardwell JC (2008) Disulfide-linked protein folding pathways. Annu Rev Cell Dev Biol 24: 211–235.
33. Kadokura H, Katzen F, Beckwith J (2003) Protein disulfide bond formation in prokaryotes. Annu Rev Biochem 72: 111–135.
34. Kallis GB, Holmgren A (1980) Differential reactivity of the functional sulfhydryl groups of cysteine-32 and cysteine-35 present in the reduced form of thioredoxin from *Escherichia coli*. J Biol Chem 255: 10261–10266.

35. Holmgren A (1995) Thioredoxin structure and mechanism: conformational changes on oxidation of the active-site sulfhydryls to a disulfide. Structure 3: 239–243.
36. Hatahet F, Ruddock LW (2009) Protein disulfide isomerase: a critical evaluation of its function in disulfide bond formation. Antioxid Redox Signal 11: 2807–2850.
37. Lappi KA, Lensink MF, Alanen HI, Salo KEH, Lobell M, et al. (2004) A conserved arginine plays a role in the catalytic cycle of the protein disulfide isomerases. J Mol Biol 335: 283–295.
38. Carvalho AP, Fernandes PA, Ramos MJ (2006) Similarities and differences in the thioredoxin superfamily. Progress Biophys Mol Biol 91: 229–248.
39. Cheng Z, Zhang J, Ballou DP, Williams CH (2011) Reactivity of thioredoxin as a protein thiol-disulfide oxidoreductase. Chem Rev 111: 5768–5783.
40. Gilbert HF (1997) Protein disulfide isomerase and assisted protein folding. J Biol Chem 272: 29399–29402.
41. Holmgren A (1981) Thioredoxin: structure and functions. TIBS 6: 26–29.
42. Pigiet VP, Schuster BJ (1986) Thioredoxin-catalyzed refolding of disulfide-containing proteins. Proc Natl Acad Sci U S A 83: 7643–7647.
43. Lundström J, Krause G, Holmgren A (1992) A Pro to His mutation in active site of thioredoxin increases its disulfide-isomerase activity 10-fold. New refolding systems for reduced or randomly oxidized ribonuclease. J Biol Chem 267: 9047–9052.
44. Livingstone D (2009) A practical guide to scientific data analysis. Wiley. p. 341.
45. Röthlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, et al. (2008) Kemp elimination catalysis by computational design. Nature 453: 190–195.
46. Kozlov G, Määttänen P, Thomas DY, Gehring K (2010) A structural overview of the PDI family of proteins. FEBS J 277: 3924–3936.
47. Darby NJ, Creighton TE (1995) Characterization of the active site cysteine residues of the thioredoxin-like domains of protein disulfide isomerase. Biochemistry 34: 16770–16780.
48. Darby NJ, Penka E, Vincentelli R (1998) The multi-domain structure of protein disulfide isomerase is essential for high catalytic efficiency. J Mol Biol 276: 239–247.
49. Xiao R, Solovyov A, Gilbert HF, Holmgren A, Lündstrom-Ljung J (2001) Combinations of protein-disulfide isomerase domains show that there is little correlation between isomerase activity and wild-type growth. J Biol Chem 276: 27975–27980.
50. Wilkinson B, Gilbert HF (2004) Protein disulfide isomerase. Biochim Biophys Acta 1669: 35–44.
51. James LC, Tawfik DS (2003) Conformational diversity and protein evolution – a 60-years-old hypothesis revisited. Trends Biochem 28: 361–368.
52. Tokuriki N, Tawfik DS (2009) Protein dynamism and evolbality. Science 324: 203–207.
53. Colletier J-P, Aleksandrov A, Coquelle N, Mraihi S, Mendoza-Barberá, et al. (2012) Sampling the conformational energy landscape of a hyperthemophilic protein by engineering key substitutions. Mol Biol Evol 29: 1683–94.
54. Ben-David M, Elias M, Filippi J-J, Duñac E, Sulman I, et al. (2012) Catalytic versatility and backups in enzyme active sites: the case of serum paraoxonase 1. J Mol Biol 418: 181–96.
55. Papp B, Pal C, Hurst LD (2003) Dosage sensitivity and the evolution of gene families in yeast. Nature 424: 194–197.
56. Debarbieux L, Beckwith J (1998) The reductive enzyme thioredoxin 1 acts as an oxidant when it is exported to the *Escherichia coli* periplasm. Proc Natl Acad Sci U S A 95:10751–10756.
57. Bar-Even A, Noor E, Savir Y, Liebermeister W, Davidi D, et al. (2011) The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. Biochemistry 50: 4402–4410.
58. Perez-Jimenez R, Wiita A, Rodriguez-Larrea D, Kosuri P, Gavira JA, et al. (2008). Force-clamp spectroscopy detects residue co-evolution in enzyme catalysis. J Biol Chem 283: 27121–27129.
59. Rodriguez-Larrea D, Perez-Jimenez R, Sanchez-Romero I, Delgado-Delgado A, Fernandez J.M, et al. (2010) Role of conservative mutations in protein multi-property adaptation. Biochem J 429: 243–249.
60. Homgren A (1979) Thioredoxin catalyzed the reduction of insulin disulfides by dithiothreitol and dihydrolipoamide. J Biol Chem 254: 9627–9632.

3.3 Artículo en vías de publicación: "Constructing a scaffold for protein design via multi-approach stabilization."

Aumentar la estabilidad de las proteínas es una de las claves para desbloquear el uso de las mismas con fines biotecnológicos y terapéuticos, no solo por permitir sus aplicaciones directas, sino que también permite que sean diseñadas con el fin de optimizar una propiedad deseada o para diseño *de novo*.

La estabilidad de las proteínas ha sido aumentada por muchos métodos, el problema es que ninguno de ellos puede garantizar por sí solo, un incremento suficiente para compensar los efectos desestabilizantes asociados al diseño de nuevas funciones. Por tanto, en este trabajo hemos testado la posibilidad de combinar diferentes estrategias a fin de conseguir una proteína de gran estabilidad que pueda servir como base para el diseño molecular de nuevos sitios activos. El plan de estabilización seguido en este trabajo consiste en 3 pasos: primero estabilización por mutaciones de consenso, segundo el rediseño de la distribución superficial de carga a través de una nueva estrategia computacional-evolutiva, y por último la sustitución de un residuo enterrado ionizable por uno hidrofóbico en el centro activo.

Bajo este plan de estabilización se consiguió incrementar la estabilidad de la tiorredoxina de E. coli hasta convertirla en una de las proteínas más estables conocidas, lo que la convierte en un idóneo armazón molecular para el diseño de nuevas funciones proteicas. Estos resultados ponen por una parte de manifiesto el éxito de nuestro nuevo enfoque para el rediseño de las interacciones de carga de carga de la superficie, que consiguió aumentar la estabilidad manteniendo la actividad enzimática de la proteína. Enfoque que podría ser aplicado a estabilizar las enzimas con potencial biotecnológico. Por otro lado los resultados parecen confirmar la hipótesis de que es posible aumentar la estabilidad de una proteína combinando distintas estrategias simultáneas, lo cual podría ayudar a solventar las limitaciones asociadas a la estabilidad en muchos estudios de ingeniería de proteínas, en particular aquellos enfocados al diseño de funciones *de novo*.

Constructing a scaffold for protein design via multi-approach stabilization.

Abstract

Increasing protein stability is a keystone for the wider use of proteins for biotechnological and therapeutical purposes. A high stability not only enables proteins applications but also allows to engineering them in order to optimize a desired property or for *de novo* design. The stability of proteins has been increased by several approaches, but none can guarantee by itself an enhancement enough to compensate the potential destabilizing effects of protein design. Therefore, we tested the possibility of combining different approaches to construct a useful scaffold for protein engineering. The stabilization plan involves 3 steps: a consensus approach, the redesign of the charge surface distribution via a new computational-evolutive approach, which involves the construction and the study of a combinatorial library, and the substitution of a buried ionizable residue of the active center. This approximation enables us to increase the stability of the E coli thioredoxin converting it into one of the most stable proteins known, becoming in an appropriate scaffold for protein design. Our results suggest that it is feasible to achieve large increments in the stability of a protein by the combination of several approaches, which may have important implications in the task of constructing appropiate scaffolds for the design of new protein functions, especially for the *de novo* design.

Introduction

Designing proteins with new or improved functions very often leads to important decreases in stability, especially if it involves the construction of a new active site [1]. Therefore the design of proteins with a increased stability appears to be an unavoidable intermediate step to properly designing functions. Protein stability has been successfully increased through various approaches, ranging from the rational design [2-8], to in vitro evolution [9], including semi-rational approaches, that involves the use of evolutionary information contained in the sequence alignments [10-14]. Nevertheless in most cases these approaches have their own limitations, and the increment in stability that can be achieved by one strategy, may not be enough to compensate the destabilizing effect of the protein design. One possible solution could be to combine different strategies, but it is unclear to what extent they can be compatible. In order to check the feasibility of a protein to be stabilized simultaneously for different approaches, we propose in the present work a multi-approach optimization of E coli thioredoxin stability that involves the following 3 steps:

1. *A consensus approach*. We use as a starting point a thioredoxin variant (referred as trx*) whose stability has been previously increased by consensus engineering [14].
2. *A computational-evolutive approach to redesign the surface charge distribution*. We aimed to increase the stability of trx* by redesigning the charge-charge surface interactions. For this purpose we used a novel strategy that combines computer design with the use of evolutionary information from sequence alignments, in order to select a set of stabilizing mutations. A small combinatorial library with the mutations proposed is constructed on trx* background, and a

comparatively small number of variants were assessed by Differential Scanning Calorimetry in order to determine their thermal stability.

3. *The substitution of a buried ionizable residue of the active site by a hydrophobic residue*. It is well known that this kind of mutations usually have large stabilizing effects, in particular it is described for E. coli thioredoxin, that the substitution of a buried aspartic (D26) to a Isoleucine results in a large increase in the global stability of the protein [15]. We aimed to perform this mutation on the best library variant found in the step 2, in order to explore the possibility of converting E. coli thioredoxin in an appropriate scaffold for protein designing.

Results and Discussion

In the first step of our multi-approach stabilization we use as a starting point a E coli thioredoxin variant, Trx*, this variant was previously stabilized by consensus engineering [14]. Trx* involves the following consensus mutations on the E. coli thioredoxin (D10A, A22P, I23V, Q50A, P68A, G74S, E85Q and A87V). The second step of our stabilization plan involves the optimization of the surface charge-charge distribution of trx* in order to increase its stability. This optimization is performed by the use of a Tanford-Kirkwood based genetic algorithm [16]. The genetic algorithm (GA) procedure is used to search optimized charge-charge surface distributions, in which the charge-charge interaction energy is calculated by the Tanford-Kirkwood model that includes a correction for surface accessibility [7, 17, 18]. A number of 32 trx* positions were included in the optimization for having >50% solvent accesibility (see Figure 1A and Material and methods for more details). After running the GA, the 21 most favorable charge-charge surface distributions were selected, all of them were found to be local minima (in the sense that they cannot be improved by any other single mutation). The selected

distributions involved 22-27 substitutions with respect to trx*. In order to select a smaller set of mutations for the charge-charge surface optimization, we performed an alignment with the 21 distributions selected. The statistical free energies for the mutations proposed by de GA distributions alignment were calculated according to the following expression [19]:

$$\Delta\Delta G_{(a\rightarrow b)} = RTln\left(\frac{N_a}{N_b}\right) \quad (1)$$

Where $\Delta\Delta G_{(a\rightarrow b)}$ is the effect in stability of an a→b mutation at a given position. R is the universal gas constant, T is the "temperature" considered as a measure of evolutionary pressure on protein stability; for these calculations its value was fixed to 300K. N_a is the number of distributions with the same charge "a" at a given position and N_b the number of distributions with "b" charge at the same position.

Also, equation (1) was applied to another sequence alignment derived from BLAST search as previously described [12]. In this work it is suggested that the greater the statistical free energy calculated for a given mutation the more likely the given mutation has a stabilizing effect. Thus, the statistical free energy of the mutation a→b was calculated in the BLAST alignment.

Finally, the sum of the statistical free energies derived from the GA distributions and the statistical free energies derived from BLAST alignment was calculated (see Figure 1A). The 10 mutations with the highest summed values were selected for experimental testing (see Figure 1B)

A combinatorial library spanning the 10 mutations was constructed; it comprises 2^{10}=1024 variants. A number of 55 randomly selected variants were purified and their thermal stability was determined by DSC (see Figure 2A), the results showed a modulation range in the denaturation temperature of about 15 ^{o}C. Most of the variants increased the denaturation temperature of the trx*, involving increases up to 9 ^{o}C. The most stabilized variant of the library, referred as V7 (with the following mutations: H6E, T14K, T54K,

Q62K, Q85K, Q98E, A105K over the trx* background) showed a denaturation temperature of 114ºC.

The natural enzymatic activity (reductase activity) of V7 was measured revealing that, despite the high increase in stability achieved, it retains almost the same level of activity that trx* (see Figure 3), and a higher level if we compared with the Wt thioredoxin [20]. It is remarkable that the stability increment achieved is quite substantial especially considering that trx* is already a stabilized form of Wt thioredoxin, suggesting that indeed our approach can efficiently find protein variants with enhanced thermal stability without losing their enzymatic skills.

With the purpose of going further in the stabilization of the E. coli thioredoxin, we decided to add another strategy to the stabilization plan explained above, that involves the substitution ionizable-to-hydrophobic in a buried residue of the active site. It is well known that burial polar residues tend to have destabilizing effects in proteins. Stephen Mayo and coworkers identified that the substitution of a buried Aspartic (D26) to a hydrophobic residue such as Isoleucine results in a large increase in the global stability [15]. Thus mutation D26I was made over the V7 variant. The above library variants were prepared with a His_6 tag in order to facilitate the purification, but in despite the small contribution of the His tag to the stability observed in the present work, the variant $V7_{D26I}$ was prepared without the His_6 tag in order to perform a more detailed analysis. The denaturation temperature of $V7_{D26I}$ showed a strong scan rate dependence (see Figure 2B), and the subsequent extrapolation to the infinite scan rate give a denaturation temperature of approximately 127ºC (see Figure 2B), that is 20ºC higher than that described for the trx* without the His_6 tagg, and 40ºC than the one of Wt thioredoxin [14], becoming actually one of the most stable proteins known. The transitions obtained display a significant degree of reversibility,

calculated as the ratio of the areas enclosed between first and reheating calorimetric runs.

It is known that Aspartic 26 play an essential role in the enzymatic activity [21], so changing it for an Isoleucine is expected to practically abolish the activity. However reductase activity of the $V7_{D26I}$ variant was measured (see Figure 3) showing a significant level of catalytic activity, specifically retaining about a 20% of the trx* activity.

Under our understanding this results show on one hand that we have successfully developed a new approach for stabilizing proteins, which combines a computational tool that search for optimal charge-charge surface distributions, the information obtained from sequence alignments and the construction of a small library of variants. The methodology could be used as a quick and easy way to obtain active proteins with increased stability, and involving only a few mutations. These features make it interesting for dealing with enzymes with biotechnological applications. On the other hand, the results seem to confirm the main hypothesis of this work that is that proteins can be successfully stabilized by several simultaneous approaches, which may have important implications in the task of constructing robust scaffolds for the design of new protein functions, especially for the *de novo* design.

Materials and Methods

Sequence alignment and genetic algorithm

The sequences alignment was performed using BLAST2 (1996-2003, W. Gish http://blast.wustl.edu) to search the Swiss-Prot sequence database of May 2009 (www.ebi.ac.uk/swissprot/), the sequence of E. coli thioredoxin was used as query. The Smith-Waterman algorithm was used to align the sequences found with the query. Sequences with lower identity than 0.25 were discarded for further analysis.

The optimization of charge-charge surface interaction energies of the protein, calculate by the TKSA model [7, 17], was performed using a homemade GA [16]. Mutations of the trx* were simulated *in silico* over the X-ray structure of E. coli thioredoxin (2trx.pdb) and this simulation was used as a template for the model. The simulation was made by choosing the best rotamers as provided by the GROMOS96 implementation of Swiss-Pdb Viewer [22]. Best rotamers are considered according to the score provided by Swiss-Pdb Viewer; in case of having the same score value it was chosen the one with the lower van der Waals interaction energy. Sites chosen for introducing ionizable residues were those that are >50% solvent exposed. The trx* residues selected for the optimization were: (S1, E2, H6, A10, T14, K18, E20, E30, W31, P34, K36, M37, P40, E47, A50, K52, T54, Q62, P64, A68, K69, K73, S74, N83, Q85, V91, S95, Q98, E101, E104, A105, A108). The solvent accessibility of the residues was calculated using a modification of the Shake-Rupley algorithm. During the GA runs it was allowed to every site included in the optimization to have a positive, negative o neutral charge.

Thioredoxin variants generation and purification

The combinatorial library of thioredoxin variants on the trx* background was constructed by gene assembly mutagenesis as previously described [14]. In order to facilitate the purification, the library variants included a His_6 tag at the amino-terminal end. Expression, purification, evaluation of the purity and concentration measurements was done as described [14]. The variant $V7_{D26I}$ were prepared by site-directed mutagenesis [19] on the V7 background. A $V7_{D26I}$ version without the His_6 tagg was also prepared and purificated as described [19].

Differential scanning calorimetry (DSC).

Differential scanning calorimetry experiments were performed in a capillary VP-DSC microcalorimeter (MicroCal, General Electric). Protein solutions were

extensively dialyzed against 50 mM Hepes pH:7. Experiments of the library variants were performed at a scan rate of 2.5 K/min. The experiments for the $V7_{D26I}$ variant were performed at 4, 2, 1, 0.5 K/min. All DSC experiments were carried out at protein concentrations about 0.5mg/ml or bellow. Reheating runs were performed in order to check for reversibility.

Activity messurements.

Reductase activity was determined using the turbidimetric assay of the thioredoxin catalyzed reduction of insulin [23] as previously described [20].

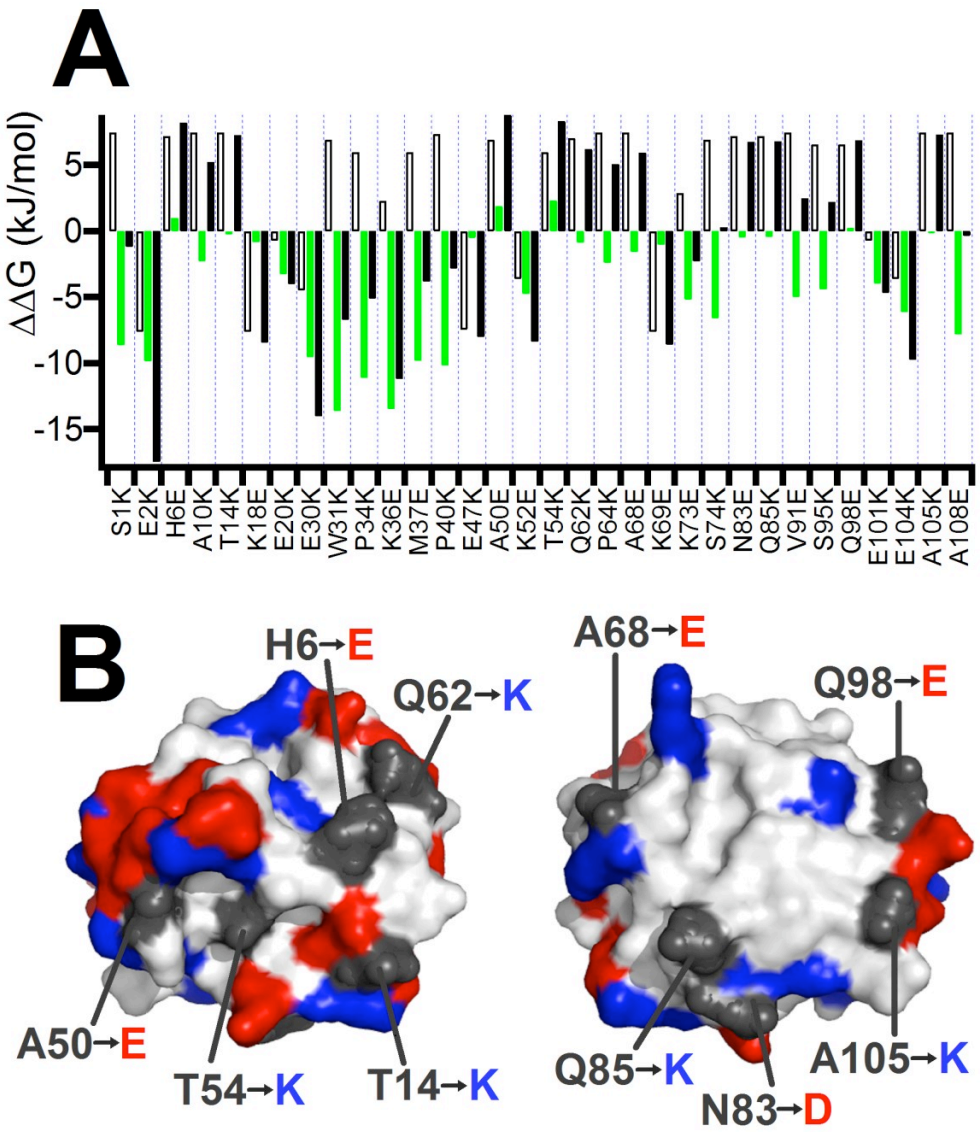

Figure 1. **A** Statistical free energies calculated according to equation (1), for the mutants proposed by the GA distributions. The mutations represented in the x axis were ones that showed the higher statistical free energies in the GA distributions alignment. Empty bars represents the statistical free energies for the mutations in the GA distributions alignment, solid green bars represents the statistical free energies of the mutations in the BLAST sequences alignment, and in solid black bars are represented the sum of both statistical energies. **B** Surface representation of the E. coli thioredoxin strcuture (2trx). In blue are shown the positive charged residues, in red the negative charged residues, and in dark grey the 10 residues selected for mutation, corresponding mutations are labeled.

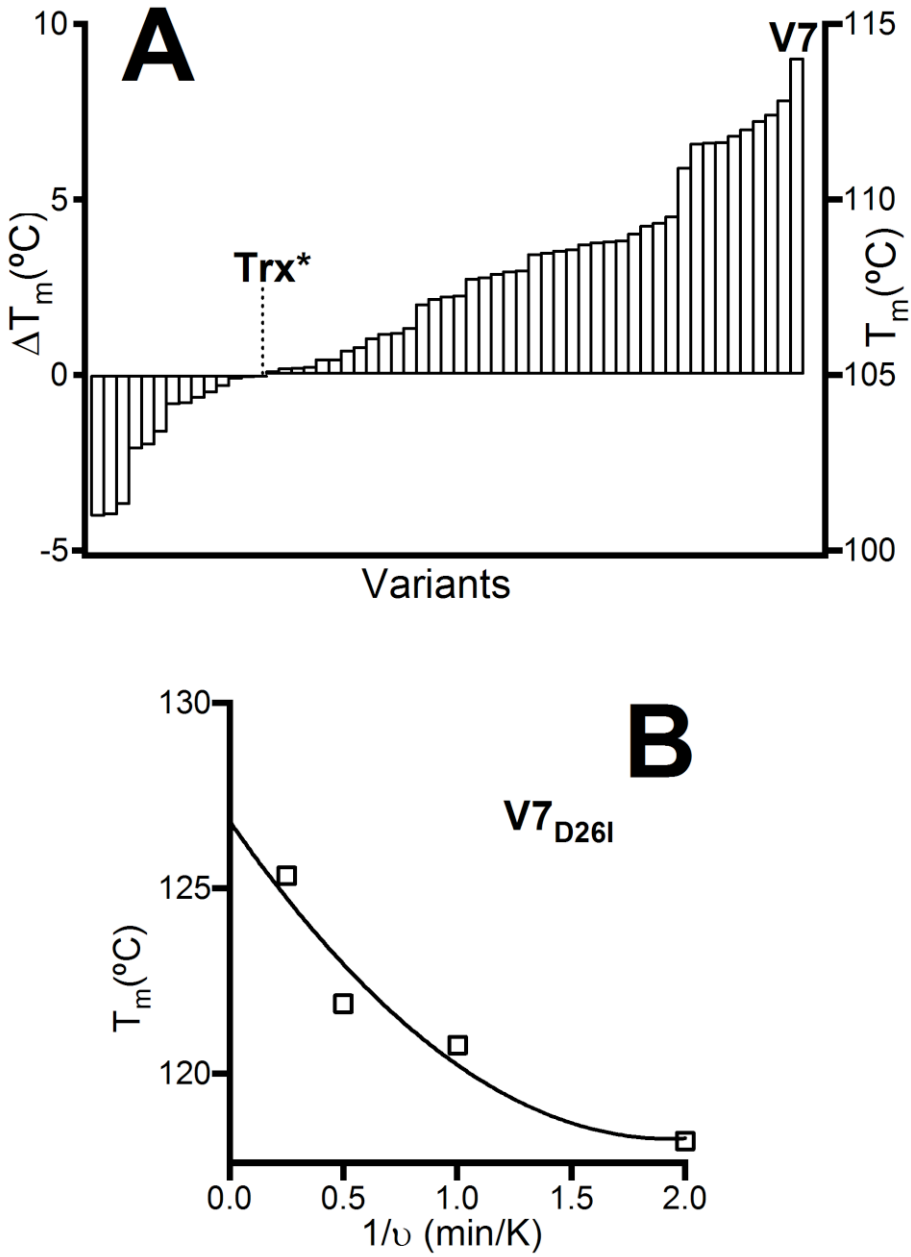

Figure 2. Denaturation temperature plots of the thioredoxin variants studied. **A** Denaturation temperature increment with respect to trx* for the library variants studied. trx* and most stabilized variant (V7) are labeled. In the right axis it is shown the denaturation temperature scale. **B** Denaturation temperature of the variant $V7_{D26I}$ versus the inverse of DSC scan rate. The data were fitted to a second-degree polynomial. The extrapolation to infinite scan rate gives a real value of T_m of 127ºC.

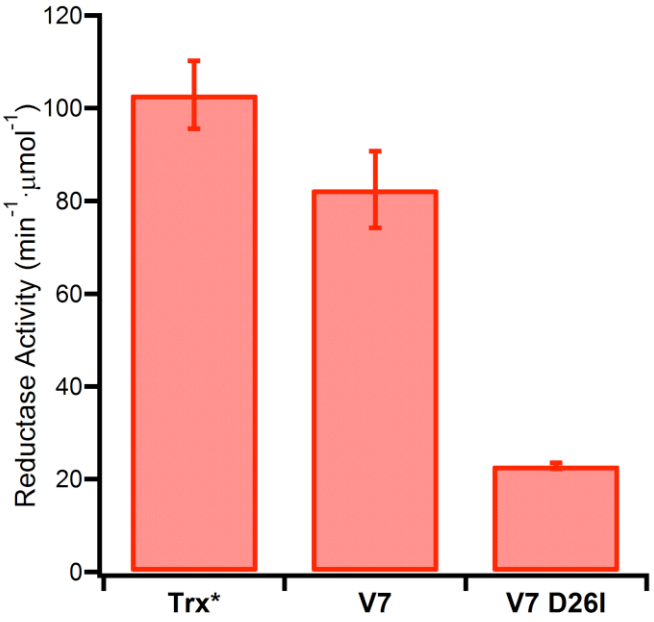

Figure 3. Reductase activities of Trx*, the best library variant V7 and $V7_{D26I.}$ The activities were determined at 37°C by a turbidimetric assay of the thioredoxin catalyzed rate of reduction of insulin [23].

References

1. Tokuriki, N., et al., *How protein stability and new functions trade off.* PLoS Comput Biol, 2008. **4**(2): p. e1000002.
2. Eijsink, V.G., et al., *Rational engineering of enzyme stability.* J Biotechnol, 2004. **113**(1-3): p. 105-20.
3. Korkegian, A., et al., *Computational thermostabilization of an enzyme.* Science, 2005. **308**(5723): p. 857-60.
4. Malakauskas, S.M. and S.L. Mayo, *Design, structure and stability of a hyperthermophilic protein variant.* Nat Struct Biol, 1998. **5**(6): p. 470-5.
5. Fu, H., et al., *Increasing protein stability by improving beta-turns.* Proteins, 2009. **77**(3): p. 491-8.

6. Gribenko, A.V., et al., *Rational stabilization of enzymes by computational redesign of surface charge-charge interactions.* Proc Natl Acad Sci U S A, 2009. **106**(8): p. 2601-6.
7. Loladze, V.V., et al., *Engineering a thermostable protein via optimization of charge-charge interactions on the protein surface.* Biochemistry, 1999. **38**(50): p. 16419-23.
8. Sanchez-Ruiz, J.M. and G.I. Makhatadze, *To charge or not to charge?* Trends Biotechnol, 2001. **19**(4): p. 132-5.
9. Eijsink, V.G., et al., *Directed evolution of enzyme stability.* Biomol Eng, 2005. **22**(1-3): p. 21-30.
10. Lehmann, M., et al., *The consensus concept for thermostability engineering of proteins: further proof of concept.* Protein Eng, 2002. **15**(5): p. 403-11.
11. Lehmann, M., et al., *The consensus concept for thermostability engineering of proteins.* Biochim Biophys Acta, 2000. **1543**(2): p. 408-415.
12. Godoy-Ruiz, R., et al., *A stability pattern of protein hydrophobic mutations that reflects evolutionary structural optimization.* Biophys J, 2005. **89**(5): p. 3320-31.
13. Godoy-Ruiz, R., et al., *Relation between protein stability, evolution and structure, as probed by carboxylic acid mutations.* J Mol Biol, 2004. **336**(2): p. 313-8.
14. Rodriguez-Larrea, D., et al., *Role of conservative mutations in protein multi-property adaptation.* Biochem. J., 2010. **429**(2): p. 243-9.
15. Bolon, D.N., et al., *Prudent modeling of core polar residues in computational protein design.* J Mol Biol, 2003. **329**(3): p. 611-22.
16. Ibarra-Molero, B., Sanchez-Ruiz, J.M., *Genetic algorithm to design stabilizing surface-carge distributions in proteins.* J Phys Chem B, 2002. **106**: p. 6609-6613.

17. Ibarra-Molero, B., et al., *Thermal versus guanidine-induced unfolding of ubiquitin. An analysis in terms of the contributions from charge-charge interactions to protein stability.* Biochemistry, 1999. **38**(25): p. 8138-49.

18. Matthew, J.B. and F.R. Gurd, *Calculation of electrostatic interactions in proteins.* Methods Enzymol, 1986. **130**: p. 413-36.

19. Godoy-Ruiz, R., et al., *Natural selection for kinetic stability is a likely origin of correlations between mutational effects on protein energetics and frequencies of amino acid occurrences in sequence alignments.* J Mol Biol, 2006. **362**(5): p. 966-78.

20. Garcia-seisdedos, H., Ibarra-Molero, B., Sanchez-Ruiz, J.M., *Probing the mutational interplay between primary and promiscuous protein functions: A computational-experimental approach.* PLoS Computational Biology, 2012.

21. LeMaster, D.M., P.A. Springer, and C.J. Unkefer, *The role of the buried aspartate of Escherichia coli thioredoxin in the activation of the mixed disulfide intermediate.* J Biol Chem, 1997. **272**(48): p. 29998-30001.

22. van Gunsteren, W.F., Billeter, S.R., Eising, A.A., Hünenberger, P.H., Krüger, P., Mark, A.E., Scott, W.R.P., Tironi, I.G., *Biomolecular Simulation: The GROMOS96 Manual and User Guide.* VdF: Hochschulverlag AG an der ETH Zürich and BIOMOS b. v, Zürich, Groningen, 1996; ISBN 3 7281 2422 2.

23. Holmgren, A., *Thioredoxin catalyzes the reduction of insulin disulfides by dithiothreitol and dihydrolipoamide.* J Biol Chem, 1979. **254**(19): p. 9627-32.

4 Resumen y Conclusiones

En la presente Tesis hemos abordado estudios relacionados con objetivos clave en la ingeniería de proteínas, como son el diseño de sitios activos, el aumento de la termoestabilidad y el aumento de actividades promiscuas. Dichos estudios fueron llevados a cabo en base al análisis de pequeñas bibliotecas combinatoriales de variantes proteicas, cuyas mutaciones fueron propuestas por distintos criterios en cada caso, bien estructurales, evolutivos o computacionales, con el fin de focalizar la búsqueda de mutaciones beneficiosas. Hay dos aspectos novedosos en la metodología utilizada que conviene resaltar. En primer lugar, la aplicación de una técnica de regresión multivariante, la llamada regresión por mínimos cuadrados parciales, comúnmente utilizada en las ciencias sociales, que nos puede permite, con el estudio de unas pocas variantes, modelar la relación entre genotipo y fenotipo, y de esta forma inferir el comportamiento de toda la biblioteca, y por ende acceder a las mejores variantes, incluso en casos en los que el cribado de alto rendimiento no sea posible. En segundo lugar, y relacionado con objetivos de optimización simultanea de varias características, hemos utilizado un enfoque basado en la predicción de la frontera de Pareto, un concepto utilizado comúnmente en economía y que se ha empezado a utilizar muy recientemente en ingeniería de proteínas.

La eficacia de estas metodologías ha sido testada en esta Tesis con bibliotecas combinatoriales pequeñas, aunque creemos que pueden convertirse por sus características en herramientas provechosas para el estudio de grandes bibliotecas de mutantes.

Respecto a los objetivos específicos abordados, cabe reseñar en primer lugar que hemos demostrado la posibilidad de combinar procedimientos de estabilización electrostáticos y evolutivos para conseguir proteínas de muy alta estabilidad. Este resultado es de gran importancia general dado que la estabilidad es una limitación en muchos estudios de ingeniería de proteínas, en particular aquellos enfocados al diseño de funciones *de novo*.

En segundo lugar, hemos demostrado que proteínas estabilizadas pueden ser suficientemente robustas como para soportar muchas mutaciones disruptivas de

tipo residuo hidrofóbico a residuo ionizable, incluso cuando se introducen en la misma región de la estructura. Este resultado puede ser relevante para la ingeniería de sitios activos y puede tener implicaciones para la compresión de la evolución de las primeras enzimas.

Finalmente, nos hemos planteado desarrollar una metodología evolutiva general para el estudio de la modulación simultánea de las actividades promiscua y nativa en proteínas. La aplicación de esta metodología al origen de la catálisis del plegamiento en dominios de tiorredoxina revela un escenario inesperado: diversos patrones de modulación de las actividades primaria y promiscua son posibles incluyendo el aumento simultáneo en ambas. Hemos demostrado además que este escenario puede explicarse fácilmente en términos de la llamada hipótesis de la diversidad conformacional. En conjunto estos resultados pueden contribuir a clarificar los mecanismos de evolución de nuevas funciones. Desde otro punto de vista, la reconstrucción por mínimos cuadrados parciales, combinada con la predicción de la frontera de Pareto que hemos usado, puede constituir la base computacional de un protocolo de evolución dirigida enfocado a la mejora simultánea de varias propiedades de una proteína y creemos, por tanto, que puede abrir nuevas posibilidades en la ingeniería de proteínas multifuncionales.

5 Bibliografía

1. Kraut, D.A., K.S. Carroll, and D. Herschlag, *Challenges in enzyme mechanism and energetics.* Annu Rev Biochem, 2003. **72**: p. 517-71.
2. Schramm, V.L., *Enzymatic transition states and transition state analog design.* Annu Rev Biochem, 1998. **67**: p. 693-720.
3. Korkegian, A., et al., *Computational thermostabilization of an enzyme.* Science, 2005. **308**(5723): p. 857-60.
4. Ashworth, J., et al., *Computational redesign of endonuclease DNA binding and cleavage specificity.* Nature, 2006. **441**(7093): p. 656-9.
5. Bolon, D.N. and S.L. Mayo, *Enzyme-like proteins by computational design.* Proc Natl Acad Sci U S A, 2001. **98**(25): p. 14274-9.
6. Jiang, L., et al., *De novo computational design of retro-aldol enzymes.* Science, 2008. **319**(5868): p. 1387-91.
7. Korendovych, I.V., et al., *Design of a switchable eliminase.* Proc Natl Acad Sci U S A, 2011. **108**(17): p. 6823-7.
8. Rothlisberger, D., et al., *Kemp elimination catalysts by computational enzyme design.* Nature, 2008. **453**(7192): p. 190-5.
9. Suarez, M., et al., *Using multi-objective computational design to extend protein promiscuity.* Biophys Chem, 2010. **147**(1-2): p. 13-9.
10. Thyme, S.B., et al., *Exploitation of binding energy for catalysis and design.* Nature, 2009. **461**(7268): p. 1300-4.
11. Siegel, J.B., et al., *Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction.* Science, 2010. **329**(5989): p. 309-13.
12. Kuhlman, B., et al., *Design of a novel globular protein fold with atomic-level accuracy.* Science, 2003. **302**(5649): p. 1364-8.
13. Bloom, J.D. and F.H. Arnold, *In the light of directed evolution: pathways of adaptive protein evolution.* Proc Natl Acad Sci U S A, 2009. **106 Suppl 1**: p. 9995-10000.
14. Bloom, J.D., et al., *Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution.* Biol Direct, 2007. **2**: p. 17.

15. Gerlt, J.A. and P.C. Babbitt, *Enzyme (re)design: lessons from natural evolution and computation.* Curr Opin Chem Biol, 2009. **13**(1): p. 10-8.
16. Kazlauskas, R.J., *Enhancing catalytic promiscuity for biocatalysis.* Curr Opin Chem Biol, 2005. **9**(2): p. 195-201.
17. Nobeli, I., A.D. Favia, and J.M. Thornton, *Protein promiscuity and its implications for biotechnology.* Nat Biotechnol, 2009. **27**(2): p. 157-67.
18. Toscano, M.D., K.J. Woycechowsky, and D. Hilvert, *Minimalist active-site redesign: teaching old enzymes new tricks.* Angew Chem Int Ed Engl, 2007. **46**(18): p. 3212-36.
19. Williams, G.J., C. Zhang, and J.S. Thorson, *Expanding the promiscuity of a natural-product glycosyltransferase by directed evolution.* Nat Chem Biol, 2007. **3**(10): p. 657-62.
20. Yoshikuni, Y., T.E. Ferrin, and J.D. Keasling, *Designed divergent evolution of enzyme function.* Nature, 2006. **440**(7087): p. 1078-82.
21. Khersonsky, O., C. Roodveldt, and D.S. Tawfik, *Enzyme promiscuity: evolutionary and mechanistic aspects.* Curr Opin Chem Biol, 2006. **10**(5): p. 498-508.
22. Bolon, D.N., et al., *Prudent modeling of core polar residues in computational protein design.* J Mol Biol, 2003. **329**(3): p. 611-22.
23. Ibarra-Molero, B., Sanchez-Ruiz, J.M., *Genetic algorithm to design stabilizing surface-charge distributions in proteins.* J Phys Chem B, 2002. **106**: p. 6609-6613.
24. Sanchez-Ruiz, J.M. and G.I. Makhatadze, *To charge or not to charge?* Trends Biotechnol, 2001. **19**(4): p. 132-5.
25. Schweiker, K.L., Makhatadze, G.I., *A computational approach for the rational design of stable proteíns and enzymes: Optimization of surface charge-charge interactions*, in *Methods in Enzymology*. 2009.
26. Eijsink, V.G., et al., *Rational engineering of enzyme stability.* J Biotechnol, 2004. **113**(1-3): p. 105-20.

27. Looger, L.L., et al., *Computational design of receptor and sensor proteins with novel functions.* Nature, 2003. **423**(6936): p. 185-90.
28. Eijsink, V.G., et al., *Directed evolution of enzyme stability.* Biomol Eng, 2005. **22**(1-3): p. 21-30.
29. Jackel, C., P. Kast, and D. Hilvert, *Protein design by directed evolution.* Annu Rev Biophys, 2008. **37**: p. 153-73.
30. Johannes, T.W. and H. Zhao, *Directed evolution of enzymes and biosynthetic pathways.* Curr Opin Microbiol, 2006. **9**(3): p. 261-7.
31. Khersonsky, O., et al., *Evolutionary optimization of computationally designed enzymes: Kemp eliminases of the KE07 series.* J Mol Biol, 2010. **396**(4): p. 1025-42.
32. Khersonsky, O., et al., *Optimization of the in-silico-designed kemp eliminase KE70 by computational design and directed evolution.* J Mol Biol, 2011. **407**(3): p. 391-412.
33. Lutz, S., *Beyond directed evolution--semi-rational protein engineering and design.* Curr Opin Biotechnol, 2010. **21**(6): p. 734-43.
34. Rodriguez-Larrea, D., et al., *Role of conservative mutations in protein multi-property adaptation.* Biochem J, 2010. **429**(2): p. 243-9.
35. Shoval, O., et al., *Evolutionary Trade-Offs, Pareto Optimality, and the Geometry of Phenotype Space.* Science, 2012.
36. Suarez, M., et al., *Pareto optimization in computational protein design with multiple objectives.* J Comput Chem, 2008. **29**(16): p. 2704-11.
37. He, L., A.M. Friedman, and C. Bailey-Kellogg, *A divide-and-conquer approach to determine the Pareto frontier for optimization of protein engineering experiments.* Proteins, 2012. **80**(3): p. 790-806.
38. Fox, R.J. and G.W. Huisman, *Enzyme optimization: moving from blind evolution to statistical exploration of sequence-function space.* Trends Biotechnol, 2008. **26**(3): p. 132-8.
39. Bessette, P.H., et al., *Construction of designed protein libraries using gene assembly mutagenesis.* Methods Mol Biol, 2003. **231**: p. 29-37.

40. Stemmer, W.P., et al., *Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides.* Gene, 1995. **164**(1): p. 49-53.
41. Privalov, P.L., *Stability of proteins: small globular proteins.* Adv Protein Chem, 1979. **33**: p. 167-241.
42. Privalov, P.L., *Stability of proteins. Proteins which do not present a single cooperative system.* Adv Protein Chem, 1982. **35**: p. 1-104.
43. Privalov, P.L., *Thermodynamic problems of protein structure.* Annu Rev Biophys Biophys Chem, 1989. **18**: p. 47-69.
44. Mateo, P.L., *Differential scanning calorimetry of protein solutions.* Thermochemistry and Its Applications to Chemical and Biochemical Systems, 1984: p. 541-568.
45. Sturtevant, J.M., *Biochemical applications of differential scanning calorimetry.* Annu. Rev. Biophys Phys. Chem., 1987. **38**: p. 463-488.
46. Sanchez-Ruiz, J.M. and P.L. Mateo, *Differential scanning calorimetry of membrane proteins.* Revis Biol Celular, 1987. **11**: p. 15-45.
47. Freire, E., et al., *Calorimetrically determined dynamics of complex unfolding transitions in proteins.* Annu Rev Biophys Biophys Chem, 1990. **19**: p. 159-88.
48. Sanchez-Ruiz, J.M., *Differential scanning calorimetry of proteins.* Subcell Biochem, 1995. **24**: p. 133-76.
49. Plotnikov, V., et al., *An autosampling differential scanning calorimeter instrument for studying molecular interactions.* Assay Drug Dev Technol, 2002. **1**(1 Pt 1): p. 83-90.
50. Plotnikov, V.V., et al., *A new ultrasensitive scanning calorimeter.* Anal Biochem, 1997. **250**(2): p. 237-44.
51. Ibarra-Molero, B.a.J.M.S.-r., *Differential scanning calorimetry of proteins: an overview and some recent advances.* In "Advanced Techniques in Biophysics" (Arrondo JL and Alonso A, eds). Elsevier, 27-48., 2006.
52. Freire, E., *Differential scanning calorimetry.* Methods Mol Biol, 1995. **40**: p. 191-218.

53. Becktel, W.J. and J.A. Schellman, *Protein stability curves.* Biopolymers, 1987. **26**(11): p. 1859-77.

54. Schellman, J.A., *The thermodynamic stability of proteins.* Annu Rev Biophys Biophys Chem, 1987. **16**: p. 115-37.

55. Privalov, P.L., et al., *Cold denaturation of myoglobin.* J Mol Biol, 1986. **190**(3): p. 487-98.

56. Woody, R., *Circular Dichroism and the conformational analysis of biomolecules.* 1996, New York: Fashman Plenum. 25-67.

57. Michl, J.a.T., E.W. , *Sprectroscopy with polarized light: Solute Alignment with Photoselection, in Liquid Crystals, Polymers, and Membranes.* 1986, New York: VCH Plublishers.

58. Kliger, D.S., Lewis, J.W., and Randall, C.E., *Polarized Light in Optics and Spectroscopy*. 1990, Academy Press: New York.

59. Lowry, T.M., *Optical Rotatory Power*. 1935, London: Green.

60. Kelly, S.M., Jess, T.J. and Price, N.C, *How to study proteins by circular dichroism.* Biochimica et Biophysica Acta (BBA) - Proteins & Proteomics, 2005. **1751**(2): p. 119-139.

61. Powis, G., Montfort, W.R., *Properties and biological activities of tioredoxins.* Annu Rev Pharmacol Toxicol, 2001. **41**: p. 261-95.

62. Holmgren, A., *Thioredoxin catalyzes the reduction of insulin disulfides by dithiothreitol and dihydrolipoamide.* J Biol Chem, 1979. **254**(19): p. 9627-32.

63. Ferrari, D.M. and H.D. Soling, *The protein disulphide-isomerase family: unravelling a string of folds.* Biochem J, 1999. **339 (Pt 1)**: p. 1-10.

64. Lundstrom, J., G. Krause, and A. Holmgren, *A Pro to His mutation in active site of thioredoxin increases its disulfide-isomerase activity 10-fold. New refolding systems for reduced or randomly oxidized ribonuclease.* J Biol Chem, 1992. **267**(13): p. 9047-52.

65. Anfinsen, C.B., et al., *The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain.* Proc Natl Acad Sci U S A, 1961. **47**: p. 1309-14.

66. Anfinsen, C.B. and H.A. Scheraga, *Experimental and theoretical aspects of protein folding.* Adv Protein Chem, 1975. **29**: p. 205-300.
67. Crook, E.M., A.P. Mathias, and B.R. Rabin, *Spectrophotometric assay of bovine pancreatic ribonuclease by the use of cytidine 2':3'-phosphate.* Biochem J, 1960. **74**: p. 234-8.
68. Hillson, D.A., N. Lambert, and R.B. Freedman, *Formation and isomerization of disulfide bonds in proteins: protein disulfide-isomerase.* Methods Enzymol, 1984. **107**: p. 281-94.
69. Ellman, G.L., *Tissue sulfhydryl groups.* Arch Biochem Biophys, 1959. **82**(1): p. 70-7.
70. McLellan, T., *Electrophoresis buffers for polyacrylamide gels at various pH.* Anal Biochem, 1982. **126**(1): p. 94-9.
71. Geladi, P., Kowalski, B.R., *Partial Least-Squares Regression: A tutorial* Analytica Chimica Acta, 1986. **185**: p. 1-17.
72. Naes, T., Isaksson, T., Fearn, T. , Davies, T., *A user-friendly guide to multivariate calibration and classification.* 2004, Chichester, UK: NIR Publications.
73. Wold, H., *Estimation of principal components and related models by iterative least squares*. Multivariate Analysis, ed. I.P.R. Krishnaiaah. 1966, New York: Academic Press.
74. Wold, S., Sjöström, M., Eriksson, L., *PLS-regression: a basic tool of chemometrics.* Chemometrics and Intelligent Laboratory Systems, 2001. **58**: p. 109-130.
75. Esbensen, K.H., *Multivariate Data Analysis: in practice. An introduction to Multivariate Data Analysis and Experimental Design. 5th ed.* 2001, Oslo: CAMO AS Publ.
76. Martens, H., Naes, T., *Multivariate Calibration.* 1989, New York: Jonh Wiley & Sons Limited.
77. Fox, R., et al., *Optimizing the search algorithm for protein engineering by directed evolution.* Protein Eng, 2003. **16**(8): p. 589-97.

78. Fox, R., *Directed molecular evolution by machine learning and the influence of nonlinear interactions.* J Theor Biol, 2005. **234**(2): p. 187-99.
79. Fox, R.J., et al., *Improving catalytic function by ProSAR-driven enzyme evolution.* Nat Biotechnol, 2007. **25**(3): p. 338-44.
80. Efron, B., *Bootstrap Methods : another look at the Jackknife.* 1979.
81. Manly, F.J., *Randomization, Bootstrap and Monte Carlo methods in Biology.* 1997, London, UK: Chapman & Hall.
82. Smith, T.F. and M.S. Waterman, *Identification of common molecular subsequences.* J Mol Biol, 1981. **147**(1): p. 195-7.
83. Dokholyan, N.V. and E.I. Shakhnovich, *Understanding hierarchical protein evolution from first principles.* J Mol Biol, 2001. **312**(1): p. 289-307.
84. Godoy-Ruiz, R., et al., *A stability pattern of protein hydrophobic mutations that reflects evolutionary structural optimization.* Biophys J, 2005. **89**(5): p. 3320-31.
85. Godoy-Ruiz, R., et al., *Relation between protein stability, evolution and structure, as probed by carboxylic acid mutations.* J Mol Biol, 2004. **336**(2): p. 313-8.
86. Shortle, D., *Propensities, probabilities, and the Boltzmann hypothesis.* Protein Sci, 2003. **12**(6): p. 1298-302.
87. Jain, R.K. and R. Ranganathan, *Local complexity of amino acid interactions in a protein core.* Proc Natl Acad Sci U S A, 2004. **101**(1): p. 111-6.
88. Ferguson, A.D., et al., *Signal transduction pathway of TonB-dependent transporters.* Proc Natl Acad Sci U S A, 2007. **104**(2): p. 513-8.
89. Halabi, N., et al., *Protein sectors: evolutionary units of three-dimensional structure.* Cell, 2009. **138**(4): p. 774-86.
90. Hatley, M.E., et al., *Allosteric determinants in guanine nucleotide-binding proteins.* Proc Natl Acad Sci U S A, 2003. **100**(24): p. 14445-50.
91. Lee, J., et al., *Surface sites for engineering allosteric control in proteins.* Science, 2008. **322**(5900): p. 438-42.

92. Shulman, A.I., et al., *Structural determinants of allosteric ligand activation in RXR heterodimers.* Cell, 2004. **116**(3): p. 417-29.
93. Reynolds, K.A., R.N. McLaughlin, and R. Ranganathan, *Hot spots for allosteric regulation on protein surfaces.* Cell, 2011. **147**(7): p. 1564-75.
94. Lockless, S.W. and R. Ranganathan, *Evolutionarily conserved pathways of energetic connectivity in protein families.* Science, 1999. **286**(5438): p. 295-9.
95. Perez-Jimenez, R., et al., *A simple tool to explore the distance distribution of correlated mutations in proteins.* Biophys Chem, 2006. **119**(3): p. 240-6.
96. Ibarra-Molero, B., et al., *Thermal versus guanidine-induced unfolding of ubiquitin. An analysis in terms of the contributions from charge-charge interactions to protein stability.* Biochemistry, 1999. **38**(25): p. 8138-49.
97. Schweiker, K.L. and G.I. Makhatadze, *A computational approach for the rational design of stable proteins and enzymes: optimization of surface charge-charge interactions.* Methods Enzymol, 2009. **454**: p. 175-211.
98. Tanford, C., Kirkwood, J.G., *Theory of protein titration curves. I. General equations for impenetrable spheres.* J. Am. Chem. Soc., 1957. **79**: p. 5333-5339.
99. Matthew, J.B. and F.R. Gurd, *Stabilization and destabilization of protein structure by charge interactions.* Methods Enzymol, 1986. **130**: p. 437-53.
100. Matthew, J.B. and F.R. Gurd, *Calculation of electrostatic interactions in proteins.* Methods Enzymol, 1986. **130**: p. 413-36.
101. Tanford, C. and R. Roxby, *Interpretation of protein titration curves. Application to lysozyme.* Biochemistry, 1972. **11**(11): p. 2192-8.
102. Godoy-Ruiz, R., et al., *Empirical parametrization of pK values for carboxylic acids in proteins using a genetic algorithm.* Biophys Chem, 2005. **115**(2-3): p. 263-6.
103. Gribenko, A.V., et al., *Rational stabilization of enzymes by computational redesign of surface charge-charge interactions.* Proc Natl Acad Sci U S A, 2009. **106**(8): p. 2601-6.
104. Strickler, S.S., et al., *Protein stability and surface electrostatics: a charged relationship.* Biochemistry, 2006. **45**(9): p. 2761-6.

105. Isom, D.G., et al., *High tolerance for ionizable residues in the hydrophobic interior of proteins.* Proc Natl Acad Sci U S A, 2008. **105**(46): p. 17784-8.

106. Isom, D.G., et al., *Charges in the hydrophobic interior of proteins.* Proc Natl Acad Sci U S A, 2010. **107**(37): p. 16096-100.

107. Harris, T.K. and G.J. Turner, *Structural basis of perturbed pKa values of catalytic groups in enzyme active sites.* IUBMB Life, 2002. **53**(2): p. 85-98.

108. Karp, D.A., M.R. Stahley, and B. Garcia-Moreno, *Conformational consequences of ionization of Lys, Asp, and Glu buried at position 66 in staphylococcal nuclease.* Biochemistry, 2010. **49**(19): p. 4138-46.

109. Wharsel, A., Sharma, P.K., Kato, M., Parson, W.W., *Modeling electrostatic effects in proteins.* Biochim Biophys Acta, 2006. **1764**: p. 1647-1676.

110. Bosshard, H.R., D.N. Marti, and I. Jelesarov, *Protein stabilization by salt bridges: concepts, experimental approaches and clarification of some misunderstandings.* J Mol Recognit, 2004. **17**(1): p. 1-16.

111. Pey, A.L., et al., *Modulation of buried ionizable groups in proteins with engineered surface charge.* J Am Chem Soc, 2010. **132**(4): p. 1218-9.

112. Khersonsky, O. and D.S. Tawfik, *Enzyme promiscuity: a mechanistic and evolutionary perspective.* Annu Rev Biochem, 2010. **79**: p. 471-505.

113. Bar-Even, A., et al., *The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters.* Biochemistry, 2011. **50**(21): p. 4402-10.