



**UNIVERSIDAD NACIONAL AUTÓNOMA  
DE MÉXICO**

---

---

**FACTULDAD DE CIENCIAS**

**CRÓNICAS DE EVOLUCIÓN Y PLEGAMIENTO:  
GENERACIÓN DE UN BARRIL TIM QUIMÉRICO POR  
BARAJEO DE SUBDOMINIOS**

**T E S I S**

QUE PARA OBTENER EL TÍTULO DE:

**BIÓLOGA**

P R E S E N T A:

**ADRIANA ESPINOSA CANTÚ**

TUTOR

**DR. LORENZO PATRICK SEGOVIA FORCELLA**

**2007**



FACULTAD DE CIENCIAS  
UNAM



Universidad Nacional  
Autónoma de México

Dirección General de Bibliotecas de la UNAM

**Biblioteca Central**



**UNAM – Dirección General de Bibliotecas**  
**Tesis Digitales**  
**Restricciones de uso**

**DERECHOS RESERVADOS ©**  
**PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL**

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.



UNIVERSIDAD NACIONAL  
AUTÓNOMA DE  
MÉXICO

**ACT. MAURICIO AGUILAR GONZÁLEZ**  
Jefe de la División de Estudios Profesionales  
Facultad de Ciencias  
P r e s e n t e .

Por este medio hacemos de su conocimiento que hemos revisado el trabajo escrito titulado:

**“Crónicas de Evolución y Plegamiento: Generación de un Barril TIM Quimérico por Barajeo de Subdominios”**

realizado por **Espinosa Cantú Adriana**, con número de cuenta **403053399** quien opta por titularse en la opción de **Tesis** en la licenciatura en **Biología**. Dicho trabajo cuenta con nuestro voto aprobatorio.

Propietario Dr. José Luis Puente García

Propietario Dr. Arturo Carlos II Becerra Bracho

Tutor(a)  
Propietario Dr. Lorenzo Patrick Segovia Forcella

Suplente M. en C. Alfonso José Vilchis Peluyera

Suplente Biól. Juan Javier Díaz Mejía

**Atentamente**  
**“POR MI RAZA HABLARÁ EL ESPÍRITU”**  
Ciudad Univeritaria, D. F., a 17 de mayo del 2007  
**COORDINADOR DE LA UNIDAD DE ENSEÑANZA DE BIOLOGÍA**

**DR. ZENÓN CANO SANTANA**

FACULTAD DE CIENCIAS



UNIDAD DE ENSEÑANZA  
DE BIOLOGÍA

Señor sinodal: antes de firmar este documento, solicite al estudiante que le muestre la versión digital de su trabajo y verifique que la misma incluya todas las observaciones y correcciones que usted hizo sobre el mismo.

|               |  |
|---------------|--|
| \$1           | Revisión de Estudios                           |
| \$319         | Pergamino de Chicharrón                        |
| \$150-400     | Fotos para el título y otros trámites          |
| \$¿?          | Transportes de índole burocrático              |
| \$¿?          | Impresiones, tiempo de internet, etc.          |
| \$1,000-2,000 | Impresión de tesis                             |
| \$805         | Cédula profesional de licenciatura y posgrado* |

Titularse SÍ tiene precio. Que se soslaye el tremendo sesgo económico que existe en la educación pública, TAMBIÉN. \*Por efectos de espacio no se enumeraron todos los gastos que un estudiante debe de hacer a lo largo de su carrera, el lector deberá recapitularlos por su cuenta.

## AGRADECIMIENTOS Y DEDICATORIAS

Tuve alrededor de seis largos meses, para ser justos, para decidirme por mis dedicatorias y agradecimientos –a las cuales les he dado vueltas y vueltas yo creo que desde el día que nací– y llego unas horas antes de la impresión de mi tesis con la mente en un escabroso blanco. Razones y justificaciones hay varias, en realidad debiéndose a una mezcla de sentir que todo lo que pueda escribir sería insuficiente con una pizca de que, debido al estrés que este proceso ha generado, me siento con el deber de proteger a toda persona de verse inmiscuida en él. No se salvarán, sin embargo, algunas personas que nombraré a continuación, a las cuales les agradezco sin algún resquicio de duda todo lo que hemos y habremos de compartir y a las cuales me gustaría dedicar (temporal y sustancialmente) mucho más:

A mi madre, por supuesto. A mi papá y hermana. A mis abuelos (divinos por cierto), tíos, primos y sobrinos. Al grupo de amigos a cuya mayoría conozco desde más del 80% de mi vida y a los más nuevos, de la facultad, del instituto y de otros rinconcitos. A Ricardo Rodríguez de la Vega. A todas aquellas personas que me han abierto los ojos, oídos, mente y activado mis manos y aquello que asignamos al corazón. A Inti. A las personas que no tienen las oportunidades de dedicar su tesis de licenciatura y, sobre todo, a la gente que lucha por que las tengan. Gracias.

## AGRADECIMIENTOS

Agradezco a Lorenzo Segovia por brindarme la oportunidad de estar en su grupo de trabajo, orientarme y apoyarme.

A la gente que constituye al laboratorio por sus consejos académicos (la otra parte sentimental debe ser considerada desde lo escrito arriba): a Viviana Escobar, Areli Morán, Ernesto Perez, Javier Díaz, Alejandro Sánchez, Iliana Escamilla, Mariana Peimbert, Arcadio Farías, Haven López, Dagoberto Armenta, Mariana Buitrón, Georgina Hernández y Laura Domínguez.

A Areli del Carmen Morán García, excelente técnico académico.

A las personas encargadas de la Unidad de Síntesis y Secuenciación de ADN: Paúl Gaytán, Eugenio López, Jorge Yáñez y Santiago Becerra.

A mis sinodales: Jose Luis Puente, Arturo Becerra, Alfonso Vilchis y Javier Díaz.

# CRÓNICAS DE EVOLUCIÓN Y PLEGAMIENTO: GENERACIÓN DE UN BARRIL TIM QUIMÉRICO POR BARAJEO DE SUBDOMINIOS.

## ÍNDICE

|   |    |
|---|----|
| 1. RESUMEN  | 4  |
| 2. INTRODUCCIÓN   | 5  |
| 2.1 EVOLUCIÓN PROTEICA  | 6  |
| 2.1.1 Evolución proteica y estructura primaria  | 6  |
| 2.1.2 Restricciones evolutivas derivadas de la estructura secundaria                      | 9  |
| 2.1.3 Restricciones evolutivas derivadas de la estructura terciaria                       | 11 |
| 2.2 LA INGENIERÍA DE PROTEÍNAS  | 14 |
| 2.3 EL BARRIL TIM   | 14 |
| 2.3.1 La fosforibosil antranilato isomerasa   | 17 |
| 3. ANTECEDENTES   | 18 |
| 4. OBJETIVOS  | 21 |
| 4.2 HIPÓTESIS   | 21 |
| 4.1 JUSTIFICACIÓN   | 21 |
| 5. MATERIALES Y MÉTODOS   | 23 |
| 5.1 DISEÑO DE LOS OLIGONUCLEÓTIDOS Y PCR  | 23 |
| 5.2 SISTEMA DE COMPLEMENTACIÓN DE LA PRAI <sub>Ec</sub> CON PARÁLOGOS                     | 24 |
| 5.3 OTRAS CONSIDERACIONES METODOLÓGICAS   | 28 |
| 6. RESULTADOS Y DISCUSIÓN   | 30 |
| 6.1 COMPLEMENTACIÓN DE LA PRAI <sub>Ec</sub> CON FRAGMENTOS DE<br>GENES PARÁLOGOS         | 32 |
| 6.1.1 Conclusión y perspectivas   | 35 |
| 6.2 DISEÑO Y PRUEBA DE OLIGOS DEGENERADOS PARA LA<br>AMPLIFICACIÓN DE ORTÓLOGOS DISTANTES | 37 |
| 6.2.1 Conclusión y perspectivas   | 41 |
| 7. CONCLUSIÓN   | 44 |
| 8. APÉNDICES  | 45 |
| 9. REFERENCIAS  | 47 |

## 1. RESUMEN

La reciprocidad entre la ingeniería de proteínas y la evolución permite que el área de la ingeniería, a falta de conocimiento sobre las constricciones funcionales y estructurales de las proteínas, se desarrolle por imitación al proceso natural, a la vez que los resultados de la ingeniería de proteínas aumentan el entendimiento de la evolución. La recombinación, tanto en la ingeniería de proteínas como en el proceso evolutivo, ha sido un fenómeno fundamental en la generación de variantes proteicas. Las recombinaciones homólogas aumentan la probabilidad de viabilidad del producto quimérico al disminuir la probabilidad de inclusión de mutaciones deletéreas y actúan predominantemente cambiando la especificidad al sustrato y/o la arquitectura del sitio activo. Por otra parte, las recombinaciones no homólogas entre varios genomas maximizan la exploración del área de secuencia, posibilitando la producción de nuevos plegamientos. En el presente trabajo se plantea una metodología basada en ambos tipos de recombinación para capacitar la generación de un banco de proteínas complementadas con fragmentos de homólogos distantes obtenidos a partir de una muestra ambiental de DNA. Para probar el sistema –basado en el tamizaje por cloranfenicol acetil transferasa, un reportero de plegamiento– se complementó la fosforibosil antranilato isomerasa (PRAI) de *E. coli* –de dominio de Barril TIM, a la cual se le eliminó un subdominio correspondiente al  $(\beta\alpha)_3\beta_4$ – con fragmentos de genes parálogos amplificados por PCR a partir de oligonucleótidos específicos. Se obtuvieron cuatro clonas quiméricas plegadas y estables en el citoplasma, dos de ellas con inserciones de complejos de oligonucleótidos, una con un fragmento del dominio río abajo del dominio de barril TIM y otra con un fragmento de un producto de PCR inespecífico. Por otra parte, se sintetizaron cuatro pares de oligonucleótidos degenerados, diseñados a partir de motivos y bloques –ambos representaciones de áreas conservadas dentro de un alineamiento múltiple– de cuatro superfamilias de barril TIM. Los oligonucleótidos se probaron en *E. coli* K12, genoma en el cual se logró la amplificación de un barril TIM. En comparación con la complementación por recombinación de fragmentos obtenidos a partir de DNA genómico, la complementación por fragmentos amplificados por PCR genera un acervo reducido de secuencias diferentes (el cual aumenta conforme más pares de oligonucleótidos y más genomas se incluyan en el análisis); el barajeo de homólogos distantes, sin embargo, puede aumentar las probabilidades de éxito de las recombinaciones no homólogas sin restringir la variabilidad de los productos. El método que aquí se plantea se puede aplicar en otro tipo de plegamientos para generar variantes proteicas útiles en la industria y en el entendimiento de la relación entre la función y el plegamiento.

## 2. INTRODUCCIÓN

El definir *vida* —objeto de estudio último de la Biología—, ha sido (y seguramente seguirá siendo) un gran problema desde el punto de vista epistemológico para cualquiera de las ciencias o ramas de la mismas que requieran de su planteamiento. La complicación se puede evitar, sin embargo, si se delimita su significado enlistando sus principales características, las cuales son ser susceptibles a adaptarse, a extraer energía de nutrientes, a crecer, diferenciarse y reproducirse. Asombrosamente, los requisitos para estar considerado entre lo vivo dependen casi por completo, y al menos para la vida conocida hasta hoy, de la armonía entre unos pocos fenómenos naturales y de unas cuantas moléculas, de las cuales sobresalen entre las orgánicas cuatro: carbohidratos, lípidos, ácidos nucleicos y por su gran diversidad y funcionalidad, notoriamente, las proteínas. El estudio de las proteínas y del por qué de sus características es, entonces, una estrategia para acercarnos al entendimiento de la vida misma.

La palabra proteína proviene del griego *proteion*, que significa primario o en primer lugar. Su significado adquiere sentido al considerar que, por ejemplo, las proteínas están en todo ser vivo constituyendo (a excepción de las plantas con abundante celulosa) más de la mitad de su peso seco, y que, por si fuera poco, erigen el basamento funcional y arquitectónico de la mayoría de los organelos. Estas moléculas, de las más grandes entre las biomoléculas, son cadenas de 8 hasta 100,000 aminoácidos unidos por enlaces peptídicos. Aunque se conocen más de 300 aminoácidos y teóricamente éstos podrían ser muy variados —pero siempre con la misma estructura química conformada por un átomo de carbono central unido a un grupo amino, a un grupo carboxilo, a un átomo de hidrógeno y a un grupo sustituyente—, las proteínas utilizadas por los seres vivos se construyen a partir de sólo veinte de éstos. Aún así, el número de distintas combinaciones de aminoácidos en proteínas del rango de tamaño existente resulta en un universo abrumador ( $= 20^n$ , en donde 20 son los aminoácidos comunes y  $n$  es el número de aminoácidos de la proteína) que pudiera explicar la enorme gama de funciones específicas que pueden realizar pero pareciera contradecir el limitado número, de alrededor de unas  $10^3$  a  $10^4$  variantes, de plegamientos o formas tridimensionales finales que se estima que pueden adoptar (Koonin 2002). Y es así, con esta paradoja, que surgen las preguntas subyacentes a este trabajo: ¿qué relación hay entre la secuencia, el plegamiento y la función de las proteínas?; ¿por qué existe un limitado repertorio de plegamientos que las proteínas pueden adoptar?; ¿cómo puede existir tal diversidad de funciones ante la restricción de posibles plegamientos?. Tanto como posible fenómeno participe en la explicación, como por el enfoque metodológico que su estudio confiere, la evolución ha sido esencial en la búsqueda de respuestas en el tema.

## 2.1 EVOLUCIÓN PROTEICA

La evolución de las proteínas se puede estudiar desde cada uno de sus niveles de organización, ya que cada uno confiere distintas restricciones evolutivas. Los niveles de organización de una proteína son su estructura primaria, secundaria, terciaria y cuaternaria. La estructura primaria es la secuencia ordenada de aminoácidos y determina todas las subsecuentes; la estructura secundaria es la disposición de esa secuencia de aminoácidos en el espacio, conformando  $\alpha$ -hélices u hojas- $\beta$ , agrupadas en dominios. La estructura terciaria es el total de dominios interactuando en el espacio de una forma dada, de tal manera que se genera el plegamiento proteico final. Por último, la estructura cuaternaria constituye la unión por enlaces no covalentes de varias cadenas polipeptídicas con estructura terciaria, al generar restricciones mayoritariamente en la evolución de interacciones moleculares y no en la evolución de las proteínas individuales, no será tema de estudio del presente trabajo. A continuación se presentan las características evolutivas de los primeros tres niveles de organización.

*2.1.1 Evolución proteica y estructura primaria.* Al estudiar la evolución de una proteína uno se puede que remitir hasta sus orígenes, lo cual, a su vez, requiere de la explicación del origen de las moléculas orgánicas mismas. En resumen de un interesantísimo y de igual manera extensísimo tema, existen dos hipótesis principales: la que explica que la mayoría de las moléculas orgánicas se generaron en un medio extraterrestre transfiriéndose posteriormente a la Tierra; y la que expone que la mayoría de éstas se generaron aquí mismo debido a un ambiente propicio (reductor, oxidante o neutro según la escuela) con electricidad, algún sustrato o compartimiento provechoso y a eventos repetidos de agregación, selección y evolución. Lo que sucedió posteriormente a la introducción de las moléculas orgánicas ha generado una mayor gama de hipótesis con diferencias microscópicas; una de las más aceptadas es la teoría de “el mundo del RNA”, en la cual se explica que los ácidos ribonucleicos (RNA) una vez polimerizados pudieron catalizar distintas reacciones, incluyendo la autopolimerización, y fungir, por tanto, como moléculas codificantes de sus propias cadenas y, en un periodo posterior, codificar las secuencias de aminoácidos. Partiendo de aquí, no es difícil postular, como se ha hecho formalmente, la modificación del sistema codificante dependiente del RNA al cimentado en el ácido desoxiribonucleico (DNA), más estable por naturaleza química y la forma actual más conspicua de codificación de una proteína. El sistema consta de un gen con información codificada en una secuencia de DNA que es transcrita a RNA mensajero, del cual tres nucleótidos dados –un triplete– se traducen a un aminoácido determinado. Al irse traduciendo la cadena de RNA se polimerizan los aminoácidos, formándose la estructura primaria. Ya que la cadena de DNA es la “memoria” y la “fuente informativa” para la producción de una proteína, la

evolución de las proteínas depende inicialmente de la variación que se sujeta en la secuencia de DNA.

La variación en el acervo de un gen o la diversidad alélica –en los organismos y en las poblaciones hay de una a varias copias de un gen que codifican para variantes del mismo tipo de información– se produce por mutaciones y recombinaciones. Las mutaciones pueden ser de dos tipos: puntuales o de corrimiento de pauta (Futuyma 1998). En una mutación puntual se modifica una de las bases nitrogenadas, por lo cual se puede o no modificar la estructura primaria de la proteína, dependiendo si el triplete nuevo sigue codificando para el mismo aminoácido (mutación sinónima) o para otro distinto (mutación no sinónima). Cuando hay corrimiento de pauta, la inserción, eliminación o duplicación de la(s) base(s) –pudiendo ser desde un solo nucleótido hasta toda la secuencia de un gen o varios genes– cambia la estructura primaria con mayor probabilidad. Las recombinaciones intracromosomales son otra fuente importante de diversidad alélica. Éstas pueden ser homólogas, como la recombinación que se da en la meiosis, uno de los fenómenos que hace de la reproducción sexual un éxito evolutivo, o no homólogas. En las recombinaciones homólogas se intercambia y yuxtapone información entre alelos de un mismo loci, mientras que en las no homólogas se permutan secuencias de distintos locus.

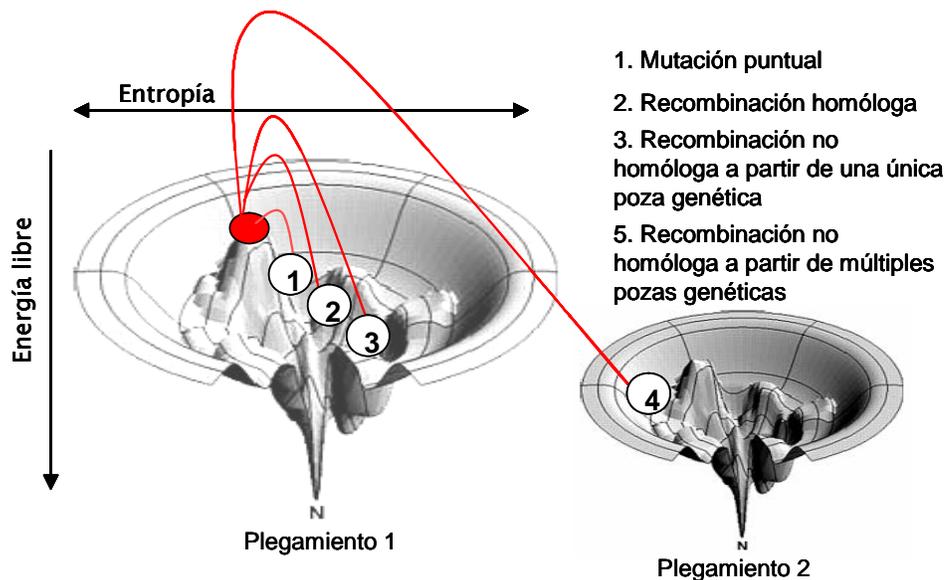
Sin importar la forma en la cual se haya iniciado la variación génica, para que haya evolución, además de la producción de variantes, se requiere que éstas se fijen, es decir, permanezcan en la población. La probabilidad de fijación depende según el caso de la adecuación de las nuevas mutantes, es decir, de la probabilidad de supervivencia y de reproducción que el nuevo carácter confiere al organismo, y de su frecuencia inicial en la población (Futuyma 1998). A grandes rasgos, en la Teoría Neutral de la Evolución se propone que las mutaciones y recombinaciones no incrementan ni disminuyen significativamente la adecuación del carácter en cuestión, por lo cual su fijación depende mayoritariamente de la deriva génica, es decir, de la frecuencia inicial del nuevo carácter, del tamaño poblacional y de la estocasticidad en la historia del linaje. La Teoría de la Selección Natural, contrariamente, se basa en que generalmente se producen variantes deletéreas o dominantes negativas y solo, muy pocas veces, variantes ventajosas, por lo cual siempre hay un cambio significativo en la adecuación (Futuyma 1998). Bajo este precepto, si el nuevo carácter es deletéreo o dominante negativo, es altamente probable que se elimine de la población, y, si es ventajoso, la selección positiva permitirá su fijación (Pal 2006).

Desde 1972, por una teoría de Susumu Ohno, se considera que la duplicación de secuencias genéticas es el factor predominante en generación de nuevos genes, ya que una vez repetida una secuencia hay menos presión selectiva, facilitándose la fijación de cualquier mutación subsiguiente a la duplicación (Patthy 2003). La duplicación es también sinónimo de la abundancia que conlleva a

la complejidad del organismo por facilitar la modularidad y la correulación (Vogel 2006); las mutaciones que le siguen, sin embargo, aunque permiten la diversificación y exploración del espacio de secuencia –posibles combinaciones de aminoácidos en una proteína–, son difícilmente trascendentes en la generación de novedades en plegamientos y funciones (Bogarad 1999). Las recombinaciones, por otro lado, han sido consideradas como el agente evolutivo principal de la versatilidad del genoma debido a que permiten una exploración más drástica del espacio de secuencia e incluso posibilitan cambios en el paisaje de adecuación –nivel de adaptación que cierto espacio de secuencia confiere, en el cual hay mínimos y máximos locales para distintos grupos de combinaciones de aminoácidos– (Bogarad 1999, Souza 2001, Vogel 2006). Dentro de los tipos de recombinaciones existen algunas diferencias en los resultados evolutivos a los que conllevan. Bogarad y Deem, en 1999 (Bogarad 1999), realizaron simulaciones matemáticas de la evolución molecular por distintos fenómenos recombinatorios que resultaron en una escala de los efectos producidos a nivel del paisaje energético. Semejante al paisaje de adaptación, el paisaje energético constituye una representación gráfica en forma de embudo en la cual el mínimo global corresponde a al estado energético más favorable que un plegamiento, a partir de una o un conjunto de secuencias puede adquirir; los mínimos y máximos locales –separados entre sí por barreras energéticas impuestas por la combinatoriedad misma– representan las distintas energías libres que resultan de otras combinaciones de secuencia, y, por tanto, distintas configuraciones que conllevan a un mismo plegamiento. De las recombinaciones, las homólogas son las menos propensas a traspasar las barreras energéticas; las recombinaciones no homólogas a partir de varias pozas génicas, las más altas en la escala, posibilitan la generación de otro paisaje energético, es decir, permiten la generación de un nuevo plegamiento (Figura 1) (Bogarad 1999).

Una vez fijada una variante empieza la divergencia. Los genes homólogos son aquellos que comparten un gen ancestral en común. Éstos pueden ser a su vez genes parálogos, ortólogos o xenólogos. Se consideran genes parálogos cuando la divergencia se produce dentro del mismo genoma por fenómenos de duplicación como por ejemplo las hemoglobinas  $\alpha$  y  $\beta$  dentro de un mismo genoma. Por otro lado, se conocen como ortólogos a los genes diferenciados por fenómenos de especiación que descienden verticalmente de un gen ancestral común y que tienden a tener la misma función. Un ejemplo de genes ortólogos son las hemoglobinas  $\beta$  en dos taxa distintos. Por último, se conocen como xenólogos a los genes homólogos que resultaron de la transferencia horizontal desde otro organismo, como lo son los genes *Fis* de enterobacterias. Como fenómeno contrario a la divergencia ocurre la convergencia, en el cual las secuencias de dos genes tienden a realizar la misma función o a codificar productos parecidos sin compartir una historia en común.

La evolución de la estructura primaria, además de representar las restricciones evolutivas iniciales de una proteína y la base de las posibles adecuaciones en los demás niveles de organización, constituye la materia prima de los análisis evolutivos por medio de la comparación de secuencias tanto nucleotídicas como de aminoácidos. Con cerca de 600 genomas completos secuenciados (KEGG 2007; Kanehisa 2006) y muchos más genes individuales, se han podido refinar tanto las historias evolutivas o filogenias de organismos como de las proteínas mismas, lo cual genera información necesaria para explicar las causas de la diversidad, del funcionamiento y de la distribución de moléculas a nivel de especies (Marsden 2006).



**Figura 1.** Efecto de procesos evolutivos en la exploración del espacio de secuencia y propensión a generar nuevos plegamientos. Ordenación jerárquica del efecto de los distintos procesos evolutivos numerados en el paisaje energético al cambiar la composición y así posibilitar cada vez en mayor grado traspasar las barreras de energía locales. Las recombinaciones no homologas a partir de múltiples pozas genéticas incluso se mueven a otro paisaje energético representante de una nueva estructura terciaria (Modificado de Dill 1997 y Bogarad 1999).

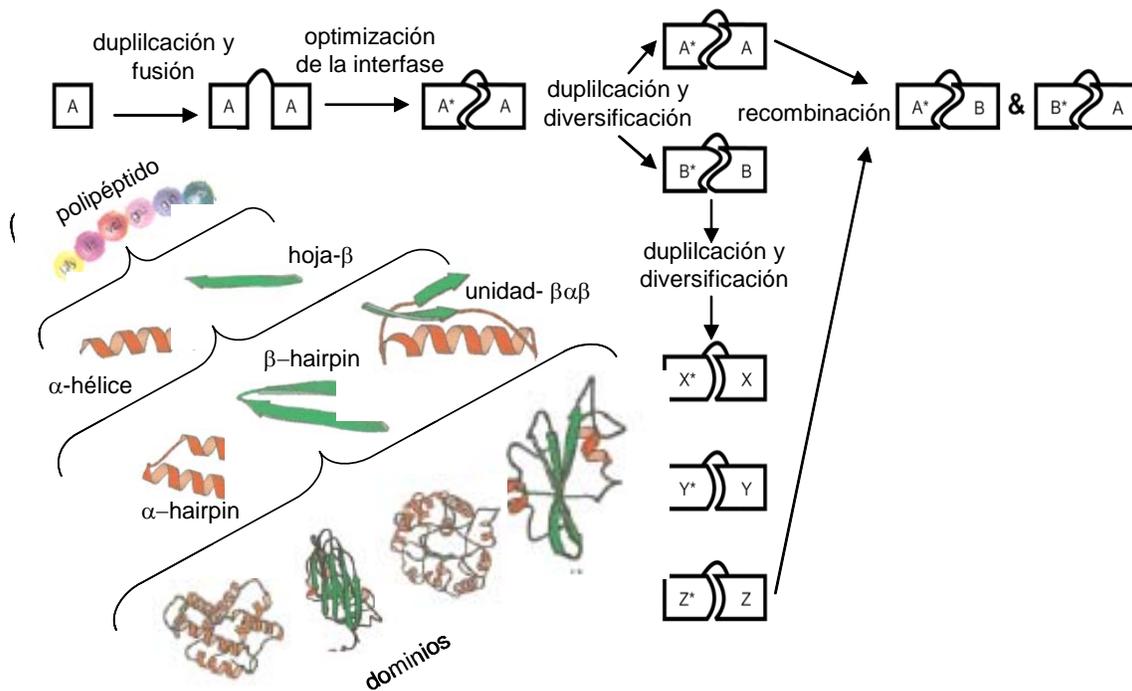
2.1.2 *Restricciones evolutivas derivadas de la estructura secundaria.* La estructura secundaria es la ordenación espacial de los grupos sustituyentes libres establecida por puentes de hidrógeno y resultante de las restricciones de los enlaces peptídicos planares a lo largo de un eje (Lehninger 1988). Los elementos más comunes de estructura secundaria son las  $\alpha$ -hélices, con un 60% de los aminoácidos en una cadena formando este tipo de estructuras, las hojas- $\beta$  y las vueltas- $\beta$ , las cuales conectan a las estructuras secundarias entre sí (Creighton 1992, Lehninger 1988). Las estructuras secundarias intervienen en el proceso evolutivo al restringir las posibilidades del plegamiento final o funcional de la proteína al ser las interacciones que se generan espontáneamente de la estructura primaria al colocar en un medio a los aminoácidos con propiedades físicas y químicas específicas. Así, si hay cambios en la estructura primaria que generen estructuras secundarias inestables, la

nueva variante será poco viable. Por otro lado, las estructuras secundarias asociadas entre sí forman estructuras supersecundarias, como el pasador- $\alpha$  ( $\alpha$ -hairpin), el pasador- $\beta$  ( $\beta$ -hairpin) y la unidad  $\beta/\alpha/\beta$ , entre otras. Las estructuras supersecundarias son unidades estables que facilitan el plegamiento y fomentan la estabilidad de la cadena polipeptídica en los subsecuentes niveles de organización, ya que promueven la regionalización de características fisicoquímicas del polipéptido, por ejemplo, la necesaria para formar el centro hidrofóbico de las proteínas globulares. A su vez, en las proteínas globulares, las estructuras supersecundarias asociadas constituyen dominios. Las definiciones de dominio varían según la disciplina que los establezca. Un dominio puede representar una unidad estable proteolíticamente, una unidad estructural, funcional, evolutiva o termodinámica (Creighton 1992, Ponting 2002). En este trabajo se utilizará dominio como parte de una cadena polipeptídica que forma una estructura compacta recurrente con más interacciones dentro de sí misma que con cualquier otra parte de la cadena (Janin 1983), sin incluir necesariamente requisitos de homología ni similitud en funcionalidad entre ejemplares de un mismo tipo de dominio. Los dominios, además de conferir estabilidad y poder dirigir el plegamiento, funcionan como bloques intercambiables en la evolución proteica, lo cual ha resultado en que se consideren como las unidades microevolutivas (Bono 2005).

Hasta el momento se han caracterizado alrededor de 93885 dominios que adoptan 40 arquitecturas distintas (CATH v.3.1.0; Orengo 1997). Su clasificación, registrada tanto en bases de datos como SCOP (Murzin 1995) y CATH, está jerarquizada de tal manera que en sus niveles más altos los dominios están definidos por la estructura tridimensional, mientras que en los más bajos por similitud de secuencia y consideraciones funcionales. La determinación de la homología entre dominios no es trivial, ya que puede haber convergencias resultantes de las ventajas adaptativas de ciertos dominios. Hoy en día existen métodos para definir si son resultantes de fenómenos convergentes o divergentes que incluyen tanto la secuencia como la estructura tridimensional y la función. El parámetro aceptado es que si un par de dominios presenta una identidad del 12% o más en posiciones estructuralmente equivalentes e igual o mayor al 35% en la secuencia, los dominios son, muy probablemente, homólogos (Ponting 2002).

Se sugiere que ciclos de selección, duplicación, conglomeración y recombinación de genes no homólogos diversificaron y dieron origen a los dominios a partir de las estructuras supersecundarias, mientras éstas se generaron, a su vez, de estructuras secundarias, y estas últimas a partir de pequeños polipéptidos (Figura 2) (Apic 2001; Pal 2006; Ponting 2002; Riechmann 2006). Los grupos de Lupas y Riechmann plantearon que después del mundo del RNA, debido a la baja fidelidad y eficiencia del sistema de transcripción, sólo se podían sintetizar pequeños polipéptidos cuyos genes eran duplicados constantemente ante la falta de un control transcripcional y que no

poseían la capacidad de autoplegarse, si no que se conformaban en el espacio según la interacción con distintas ribozimas. Posteriormente, según la misma hipótesis, se pudieron seleccionar pequeños polipéptidos que empezarían a oligomerizarse y a independizarse del RNA para conformarse espacialmente y ejercer una función. Conjunto a la aparición del DNA como templete y a sistemas de reparación, algunos genes consiguieron ensamblarse, optimizando así la termoestabilidad y el proceso de plegamiento, constituyendo las estructuras supersecundarias o los subdominos (Lupas 2001; Söding 2003; Riechmann 2006). Riechmann y Winter probaron que el ensamblaje de polipéptidos del tamaño de un subdomino sin capacidad de plegarse por sí mismos, los cuales llamaron protodomios al interactuar, pueden generar un producto estable e incluso funcional (Riechmann 2006). Experimentalmente también se ha demostrado que la recombinación no homóloga de subdominos puede generar dominios plegables con propiedades únicas (Bono 2005).



**Figura 2.** Modelo de evolución propuesto para cualquier nivel de organización proteica. Partiendo de un gen codificante para un polipéptido y tras rondas de duplicación, fusión, mutación y recombinación se puede obtener una estructura secundaria, capaz de plegarse autónomamente. El mismo proceso a partir de una estructura secundaria puede generar estructuras supersecundarias y subdominos, y estos, a su vez, dominios. La recombinación no homóloga de dominios se considera como el fenómeno más importante en la innovación de plegamientos proteicos. (La imagen del proceso evolutivo es modificada de Höcker 2002; las imágenes de los niveles de organización son modificadas de Salem 1999.)

2.1.3 *Restricciones evolutivas derivadas de la estructura terciaria.* La estructura terciaria es la conformación final de una proteína determinada por la suma de interacciones entre sus dominios, aún cuando sólo se presente uno (Bornberg-Bauer 2005). Desde que Anfinsen en 1958 estudió la

desnaturalización y renaturalización de una proteína, se sabe que la secuencia de aminoácidos (por sí sola y aunque haya chaperonas implicadas) contiene toda la información necesaria para definir la estructura tridimensional, y por tanto, su función biológica (Creighton 1992). El por qué entonces aún habiendo infinidad de secuencias distintas se calcula que solo hay de  $10^3$  a  $10^4$  plegamientos existentes y, aún más drásticamente, se estima que el 80% de todas las proteínas se conforman en tan solo 400 de estos plegamientos sobrerrepresentados, es, como se planteó anteriormente, una paradoja (Söding 2003; Koonin 2002).

Cuando ocurren mutaciones en la estructura primaria pueden suceder tres escenarios: (1) se mantiene el plegamiento original, (2) se modifica o (3) ocurre convergencia con un plegamiento preexistente. En el caso (1), en que la modificación de la estructura primaria mantiene el plegamiento original, puede no haber (1A) o haber (1B) cambios en la adecuación del mismo. Si no hay cambios en la adecuación (caso 1A), la nueva proteína estará sujeta (mayoritariamente y casi exclusivamente) a la deriva génica. Proyectando esta situación a un proceso recurrente, la explicación del por qué de la distribución de los distintos plegamientos y la razón de su poca variedad tiene que residir en el azar. Si las mutaciones en la estructura primaria generan cambios en la adecuación del plegamiento (caso 1B) por modificar su estabilidad, función, afinidad o interacción con otras moléculas, éstas estarán sujetas (no únicamente pero si evidentemente) a la selección natural. Quizás la mejor evidencia que se tiene de la selección actuando en la estabilidad *per sé* es la adaptación termófila, en la cual proteínas de organismos termófilos, con la misma función que sus homólogas en mesófilos, son mucho más estables. Las restricciones generadas por la función del plegamiento pueden ser más evidentes, ya que, una mutación en el sitio activo o en el sitio de unión a sustrato generalmente produce mutantes deletéreas. Por ello se ha concluido que los plegamientos proteicos pueden evolucionar conservando ya sea la arquitectura del sitio activo, el mecanismo de reacción y/o la especificidad del sustrato (Wise 2004). Al considerar que la selección natural puede generar constricciones importantes en la evolución de los plegamientos proteicos es plausible plantear que el limitado número de plegamientos se deba, al menos en parte, a las ventajas evolutivas que algunos de éstos confieren. Es difícil plantear las fuerzas evolutivas que resultan en los escenarios (2) y (3) debido al poco conocimiento que tenemos sobre los factores que determinan el plegamiento a partir de la estructura primaria; sin embargo el estudio de ambos es de gran importancia para entender la relación entre la estructura primaria y la terciaria, y por tanto, la paradoja anteriormente planteada.

Los mecanismos que generan cambios en las proteínas, resultando en cualquiera de estos escenarios, incluyen generalmente duplicación (seguida de mutación y selección), permutación circular y recombinación (Kinch 2002). Estos mecanismos constriñen la disyuntiva de generar

variantes con el riesgo de perder viabilidad. En el 90% de los casos de familias de proteínas con plasticidad funcional, por ejemplo, la recombinación de dominios resulta en un cambio significativo en la función de las proteínas, modificando principalmente la especificidad del sustrato y los complejos arquitectónicos funcionales (Apic 2001). Un ejemplo de la importancia del barajeo de dominios son los anticuerpos de eucariontes, cuya maduración es la forma más rápida de evolución adaptativa y la cual depende de recombinaciones y subsiguientes mutaciones. La incapacidad (energética e incluso espacial) de tener información para todos los receptores específicos para todos los antígenos posibles es superada debido a que los linfocitos que salen de la médula ósea tienen un precursor de anticuerpo generado por la recombinación al azar de segmentos de el gen VDJ. Si uno de estos precursores reconoce a un antígeno, éste es expuesto a mutaciones al azar para optimizar la afinidad, madurar y reproducirse (Loughlin 2006).

La versatilidad de combinaciones de dominios, sin embargo, está restringida (Bornberg-Bauer 2005). Partiendo del supuesto de que la duplicación de secuencias génicas y la recombinación son fenómenos azarosos, la explicación a la restricción en la combinación de dominios muy probablemente esté en la selección natural. Si un dominio o una combinación de dominios persiste, es porque es hiperadaptable (*ie.* funcional en varios nichos) o fundamental en los procesos celulares, y por tanto, propenso a la ubicuidad (Ponting 2002). La ubicuidad, por otra parte, conlleva a que estos dominios tiendan a interactuar más que aquellos menos distribuidos debido a sus capacidades adaptativas, sus propiedades topológicas o a su relevancia en las redes biológicas. El patrón de sus combinaciones también se han explicado por estocasticidad debido reflejándose en una distribución tipo *power law* o ley de potencias, en la cual se explica que hay pocos dominios muy conectados y dominios poco conectados por la alta probabilidad que existe de que los nuevos dominios se unan preferentemente a aquellos que ya están muy conectados (Koonin 2002; Vogel 2006).

Es importante puntualizar que el rearrreglo de dominios puede generar tanto divergencia como convergencia, promoviendo incluso que proteínas homólogas pueden presentar distinta estructura, lo cual dificulta su análisis evolutivo (Grishin, 2001; Ponting 2002). Sin embargo, al considerar tanto la estructura tridimensional, la secuencia de aminoácidos, la similitud de la función a nivel molecular y la organización de los dominios se puede dilucidar más fácilmente la filogenia de las proteínas y dominios. Por otro lado, el mismo análisis permite abordar el entendimiento de la relación estructura y función y dirigir las estrategias de ingeniería de proteínas a la búsqueda puntual o exhaustiva de los determinantes críticos de la estructura terciaria y la función de una proteína (Grishin 2001; Kinch 2002; Marsden 2006; Watters 2004).

## 2.2 INGENIERÍA DE PROTEÍNAS.

La naturaleza expresa solo una pequeña fracción del espacio de secuencia de todas las posibles proteínas, por lo cual buscar variantes en el laboratorio, en principio, representa una gran oportunidad; la estimación de la densidad de proteínas con función, sin embargo, es desde 1 en  $10^{12}$  hasta 1 en  $10^{77}$  variantes (Arnold 2006). Por tanto, las estrategias metodológicas en la ingeniería de proteínas tienen que optimizar la probabilidad de generar variantes propicias. Una estrategia es la evolución dirigida, en la cual se producen ciclos de evolución *in vitro* para explorar aleatoriamente el paisaje de adecuación en función de la tasa de mutagénesis impuesta sin necesidad de poseer previa información estructural o funcional detallada (Bornberg-Bauer 1999; Pal 2006; Yuan 2005). Las librerías de clonas mutantes se pueden producir por distintas técnicas –imitación de la evolución natural– generalmente por la acumulación de mutaciones secuencialmente (*eg.*, PCR mutagénica) o por recombinaciones (*eg.*, barajeo de DNA). Posteriormente al tratamiento mutagénico, las cepas se tamizan, habitualmente por resistencia a algún antibiótico, a una proteasa o por expresar la innovación de otro tipo de señal, y se puede proseguir con varios ciclos de mutagénesis y tamizaje. Otra estrategia utilizada en la ingeniería de proteínas es el diseño racional, el cual requiere de información previa de la estructura de las proteínas en cuestión, para lo cual se predicen los cambios que maximizarán la probabilidad de generar las variantes deseadas. Los logros obtenidos a partir del diseño racional se han confinado a modificaciones de la especificidad del sustrato manteniendo el mecanismo de reacción o la producción de anticuerpos catalíticos (Peimbert 2003). La aplicación del diseño racional para enfocar la evolución dirigida es, sin embargo, de gran utilidad. La evolución dirigida, en sus múltiples variantes, es hoy en día la estrategia más utilizada para la generación de actividades enzimáticas *de novo* y para la generación de hipótesis acerca de la evolución y relación entre estructura y función de las proteínas, generándose en una tasa de alrededor de 500 artículos del tema por año en los últimos tres, pero existentes desde los años 70.

## 2.3 EL BARRIL TIM

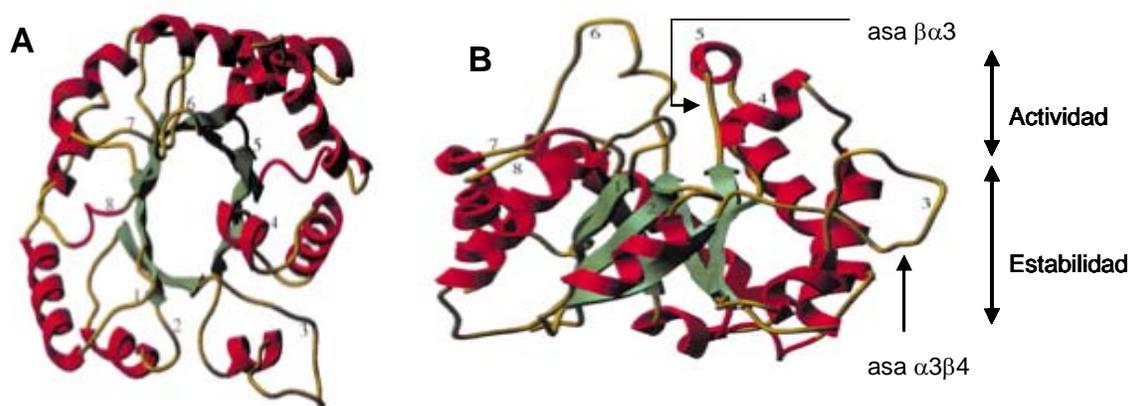
El barril TIM, plegamiento de gran atractivo en la ingeniería de proteínas, es conocido como tal por la enzima con la cual se describió el plegamiento: la triosa fosfato isomerasa (TIM), enzima dimérica encargada de interconvertir la dihidroxiacetona fosfato a gliceraldehído-3-fosfato en la glucólisis (Walden 2004; Wierenga 2001). El interés alrededor de los barriles TIM o barriles  $(\beta\alpha)_8$  en la ingeniería de proteínas se debe a algunas características inherentes a su plegamiento. El barril TIM, por ejemplo, se considera como el plegamiento más ubicuo (Wierenga 2001; Livesay 2005). Su ubicuidad en la naturaleza, suscitada por su presencia en bacterias, archeas y eucariontes, conjunto a su alto índice de diversidad dentro de los genomas y transcriptomas, permite generar

conclusiones aplicables a muchas proteínas en distintos genomas a la vez (Finke 2005; Wierenga 2001); su ubicuidad en las bases de datos estructurales, constituyendo alrededor del 10% del total de estructuras en el Protein Data Bank (PDB), permite manejar información incluyente y significativa. Por otra parte, el motivo estructural de los barriles TIM pueden ejecutar distintas funciones enzimáticas, incluyendo oxidoreductasas, transferasas, liasas, hidrolasas e isomerasas (todas las clases primarias de actividades definidas por la Comisión Enzimática menos la actividad de ligasa) (Livesay 2005). Aunado a la ventaja de su diversidad funcional, los barriles TIM se caracterizan por presentar sus sitios activos en el carboxilo terminal (C-terminal) de las hojas- $\beta$ , lo que facilita su análisis y manipulación (Soberón 2004). Por último, el barril TIM representa la paradoja dada por la diversidad en secuencias y función y la limitación estructural anteriormente planteada. El análisis filogenético entre las familias de barriles TIM a partir de las secuencias puede incluso complicarse debido a que la similitud entre las mismas se encuentra dentro de la llamada zona de penumbra o *twilight zone*, con alrededor del 20% de identidad (Chung 1996; Nagano, 2002).

El plegamiento de tipo barril TIM consiste en una estructura central formada por ocho hojas- $\beta$  paralelas rodeadas por ocho  $\alpha$ -hélices unidas entre sí mediante asas, equivalente a ocho unidades de estructuras supersecundarias conformadas por una hoja- $\beta$ , una  $\alpha$ -hélice y dos asas (Höcker 2004; Shukla 2004). Las unidades de estructuras supersecundarias suelen numerarse según su orden de aparición en la secuencia del 1 al 8 para simplificar su identificación, de tal manera que un dominio completo y normal está conformado por un barril 1-2-3-4-5-6-7-8 y la  $(\beta\alpha)_2$  se refiere a la segunda hoja- $\beta$  y a la segunda  $\alpha$ -hélice en la secuencia. En un barril  $(\beta/\alpha)_8$  cerrado, cada hoja- $\beta$  forma puentes de hidrógeno con sus dos vecinas, arrojándose de manera paralela, por lo cual se puede detectar un anillo de puentes de hidrógeno (Wierenga 2001). Su sitio activo está bien delimitado por la región de asas de los extremos carboxilo de las hojas- $\beta$ , aunque las posiciones exactas y las identidades de los residuos catalíticos pueden variar entre los barriles TIM (Figura 3) (Gerlt 2003; Livesay 2005).

La estabilidad en los barriles TIM, al contrario de la función, está comprometida hacia el amino terminal (N-terminal) de las hojas- $\beta$  (Gerlt 2003; Livesay 2005). Las unidades de estructuras supersecundarias parecen ser de plegamiento autónomo e imprescindibles en la formación de contactos entre el extremo amino y carboxilo (Finke 2005; Shukla 2004). Los barriles TIM no se pliegan de la misma manera, aunque existe una forma consenso en la cual las unidades contiguas se pliegan secuencialmente, generando intermediarios de plegamiento. Las diferencias en el proceso de plegamiento sugieren que la arquitectura proteica última no es suficiente para explicar la sucesión en el plegamiento. Las diferencias en el proceso de plegamiento pueden estar determinadas por desigualdades en la topología de contactos debido a diferentes posiciones y

tamaños de las  $\alpha$ -hélices y las hojas- $\beta$ , a los pesos energéticos de los aminoácidos constituyéndolas, a contactos no nativos (*nonnative contacts*), isomerizaciones de prolina o diferencias en la formación de puentes disulfuro (Finke 2005).

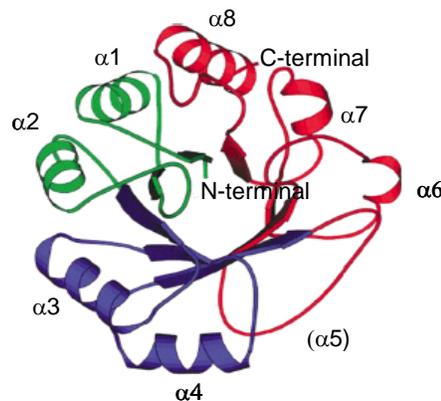


**Figura 3.** Características del plegamiento Barril TIM ilustrado en TIM de Tripanosoma (PDB 5TIM). Las  $\alpha$ -hélices están en rojo, las hojas- $\beta$  en verde. **A)** Vista superior del barril; **B)** Vista lateral. Los extremos carboxilo de las hojas- $\beta$  y las asas que conectan con las siguientes  $\alpha$ -hélices forman la cara catalítica del barril, mientras el resto es importante para la estabilidad del plegamiento. Se muestran dos asas para ejemplificar la forma en que se identifican los componentes del barril (Modificado de Höcker 2001 y Wierenga 2001).

El barril TIM generalmente es de alrededor de 250 aminoácidos, aunque existen algunos que llegan a estar constituidos únicamente por 150 residuos o hasta por 500. Las enzimas con un plegamiento de este tipo van de monoméricas a tetraméricas, como la  $\beta$ -galactosidasa, de 1023 aminoácidos (Wierenga 2001). En la base de datos CATH v3.1.0 se cuenta con 1213 estructuras clasificadas como barriles TIM, divididas en 27 superfamilias homólogas (nivel H). Las enzimas pertenecientes a una familia generalmente catalizan la misma reacción, mientras que miembros de la misma superfamilia suelen catalizar reacciones distintas pero similares en cuanto al mecanismo de acción (Livesay 2005). La diversidad funcional del plegamiento, su ubicuidad y la poca conservación a nivel de secuencia dentro de las superfamilias de barriles TIM han provocado que desde hace 20 años se discuta si son convergentes o divergentes. Los enfoques que se han utilizado, basados en la comparación de secuencias por alineamientos múltiples (Pirun 2005) y de estructuras, sugieren que la mayoría de las familias de barriles TIM surgieron por evolución divergente a partir de una proteína ancestral por reclutamiento entre vías metabólicas (Reardon 1995; Copley 2000, Livesay 2005, Pagano 2002). En un estudio realizado por Nagano *et al.* en el 2002 (Nagano 2002), se concluyó que 14 de las 17 superfamilias descritas en el 2002 tienen relaciones de identidad en secuencia, estructura y función tales que se pueden considerar como homólogas lejanas, mientras

que las 3 superfamilias restantes pueden ser homólogos extremadamente distantes o ejemplos de evolución convergente. Por otra parte, se ha planteado que las enzimas con arquitectura de barril TIM pudieron evolucionar a partir de mitades  $(\beta/\alpha)_{2N}$ , precursoras de las superfamilias, y que, por posteriores duplicaciones, mutaciones y recombinaciones formaron la subsecuente diversidad funcional y secuencial (Höcker 2004; Gerlt 2003). El entendimiento de la evolución de los barriles TIM es de suma importancia para el desarrollo de la ingeniería de estas proteínas al permitir conjuntar estrategias de diseño racional y evolución dirigida (Gerlt 2003).

2.3.1 *La fosforibosil antranilato isomerasa*. La N-(5'- Fosforibosil) Antranilato Isomerasa (PRAI) es una enzima con topología de barril TIM perteneciente a la superfamilia de Aldolasas Clase I (CATH v3.1.0) que participa en la biosíntesis de triptofano. En la mayoría de los organismos mesófilos se encuentra en estado monomérico; en *E. coli*, sin embargo, la PRAI está fusionada a la 3-glicerol-indol-fosfato sintasa (IGPS), otro barril TIM, su  $\alpha_5$  está reemplazada por un asa y presenta un giro helicoidal adicional en el asa que conecta la  $\beta_8$  con la  $\alpha_8$  (Figura 4) (Priestle 1987). El segmento  $(\beta\alpha)_{1-4}\beta_5$  de la PRAI de *E. coli* (PRAI<sub>Ec</sub>) es el más pequeño capaz de plegarse autónomamente, por lo cual se sugiere que la enzima sigue un modelo de plegamiento 5+1+2 en el cual, tras plegarse las  $(\beta\alpha)_{1-4}\beta_5$ , el fragmento es estabilizado por la unión covalente de la  $(\beta\alpha)_6$  en una primera instancia y posteriormente por las  $(\beta\alpha)_{7-8}$  (Akanuma 2005). Los mismos experimentos de fragmentación y permutación circular que permitieron concluir lo anterior demostraron que el segmento  $(\beta\alpha)_{3-4}\beta_5$  contiene interacciones esenciales en los eventos tempranos del plegamiento de la PRAI<sub>Ec</sub>, por lo cual se sugiere que este es su centro de nucleación temprana (Figura 4). Las características típicas de los barriles TIM, aunadas al tamaño pequeño de la PRAI<sub>Ec</sub>, de apenas 197 aminoácidos, soslayan sus excentricidades y la hacen un gran modelo de estudio.



**Figura 4.** Representación de listones de la estructura terciaria de la PRAI de *E. coli* en estado monomérico. Del extremo amino (N) al carboxilo (C) las hojas- $\beta$  y  $\alpha$ -hélices son enumeradas de la 1 a la 8. La  $\beta_8$  está interrumpida por un giro helicoidal adicional y la hélice- $\alpha_5$  está degenerada. Los aminoácidos del 15 al 50 se presentan en verde, son el extremo amino, implicado en el ensamblaje de la estructura final; en azul se representan los aminoácidos del 51 al 105, dentro de los cuales está el fragmento  $(\beta\alpha)_{3-4}\beta_5$ , sitio de nucleación temprana; los aminoácidos del 106 al 197 se comprenden en el fragmento rojo, extremo carboxilo que no se considera crucial para la determinación del plegamiento final (Modificado de Akanuma 2005).

### 3. ANTECEDENTES

El barajeo de DNA, técnica desarrollada por Stemmer en 1994, se consideró la revolución de la nueva era de la evolución dirigida. Imitando *in vitro* la recombinación homóloga, el barajeo de DNA exponencia la exploración del área de secuencia al generar librerías quiméricas de fragmentos de alelos o genes homólogos (Tobin 2000). La técnica consta en fragmentar al azar el conjunto de genes a barajear e hibridar los fragmentos por PCR que puede ser mutagénica para ensamblarlos en cadenas quiméricas y tamizarlas según el objetivo del experimento (Bacher 2002; Yuan 2005). Los factores limitantes en la técnica son la similitud entre los genes y el tamaño de los fragmentos. Usando modelos matemáticos del barajeo de DNA, se considera que la poza génica debe de contener homólogos con una similitud mínima del 50 al 60% (la similitud experimental que se considera para hibridar efectivamente es de 75 al 80%), incrementando la posibilidad de hibridación proporcionalmente con la similitud entre las secuencias (Moore 2001). La gran ventaja del barajeo de DNA es que se adquieren en un solo ciclo más mutaciones que por las técnicas de mutación secuencial; por otra parte, al utilizar homólogos, derivados de secuencias ancestrales que sobrevivieron a la misma presión selectiva y ser compatibles a una misma estructura, se pueden acumular más mutaciones en una secuencia sin introducir efectos deletéreos en la estructura o en la función (Drummond 2005; Yuan 2005). Desde la invención del barajeo de DNA se han desarrollado variantes de la técnica para maximizar la versatilidad de los bancos de quimeras generadas. El barajeo de módulos no homólogos (unidades funcionales, *eg.*, dominios, motivos, estructuras secundarias, *etc.*), en el cual se imita la recombinación no homóloga, es una manera de generar variantes proteicas a cualquier nivel de organización y una estrategia que ha sido utilizado satisfactoriamente en el laboratorio, superando la exploración del área de secuencia que cualquier otra técnica, hasta el momento, puede alcanzar (Yuan 2005). Mientras que con el barajeo de DNA se busca optimizar una función, con el barajeo de módulos se puede obtener variantes funcionales o en la arquitectura proteica (Yuan 2005). Desde el punto de vista metodológico, en lugar de utilizar PCRs para reensamblar los fragmentos, el barajeo de módulos se logra por ligación con enzimas, aunque otros métodos han sido desarrollados con el mismo fin (Ostermeier 1999; Tsuji 2006; Yuan 2005). En el barajeo de módulos, por tanto, no se requiere de un porcentaje de similitud determinado entre las secuencias que se utilizan.

Un precedente importante a los trabajos de barajeo es el de Doi y Yanagawa (Doi 1999), en el cual insertaron el dominio de la  $\beta$ -lactamasa TEM1 en la proteína verde fluorescente (GFP), con lo cual lograron, tras algunas rondas de PCR mutagénica, obtener variantes que tuvieran cambios de fluorescencia ante la inactivación de la  $\beta$ -lactamasa por una proteína inhibitoria, generando, así, un biosensor. Las bondades de los barriles TIM no hicieron esperar iniciativas. En 1989, Kirschner y

sus colaboradores (Luger 1989) generaron variantes del barril TIM con permutaciones circulares que resultaron prácticamente indistinguibles de la proteína inicial, mientras que en 1990 (Luger 1990) le aumentaron dos hojas- $\beta$  al extremo N-terminal de la PRAI de levadura y la expresaron en *E. coli*, en donde resultó activa y con un patrón de plegamiento similar a la nativa –concluyendo que el plegamiento tiene una plasticidad intrínseca y que éste se controla termodinámicamente tanto *in vivo* como *in vitro*–. Si bien estos tres trabajos no se basaron en el barajeo, son antecedentes importantes ya que se sustentan en el mismo principio de movilización e intercambio de módulos. Al mismo tiempo que estos trabajos se realizaban, Goraj *et al.* (Goraj 1990) utilizaban el conocimiento sobre el plegamiento de tipo barril TIM para realizar con diseño racional un subdominio  $\alpha\beta$ , el cual, imitando el proceso de evolución natural, multiplicaron para generar polipéptidos de 7, 8 y 9 unidades. El único producto con estructura compacta resultó ser el ahora conocido como *octarellin I*, de 8 unidades, tras el cual han sido generadas otras dos variantes con distintas simetrías y, por tanto, diferentes características (Goraj 1990; Houbrechts 1995). En el 2000, Riechmann y Winter (Riechmann 2000) interrumpieron el gen de la proteína de choque térmico CspA (cold shock protein), otro barril TIM, en un fragmento con aminoácidos de nucleación y lo sustituyeron con fragmentos de DNA genómico de *E. coli*, tamizando en fagos por proteólisis. Encontraron cuatro quimeras de polipéptidos solubles, una incluso con mayor estabilidad que la CspA inicial, la cual se organizó en una nueva arquitectura (Riechmann 2000; Bono 2005). En el 2003, Shukla *et al.*, (Shukla 2004) barajearon las subunidades  $\alpha\beta$  de la TIM de *S. cerevisiae* para examinar si incluso destruyendo las interacciones de hidrógeno e hidrofóbicas nativas se podía reestablecer el plegamiento, lo cual resultó ser el caso para algunas quimeras las cuales, en vez de estar constituidas por las unidades en el orden 1-2-3-4-5-6-7-8, se formaron a partir de la secuencia 2-4-6-8-1-3-5-7 y 1-3-5-7-2-4-6-8, así como de la mitad de un barril 2-4-6-8 y un cuarto de barril 1-3. Con lo anterior se propuso que el plegamiento se va generando a partir de una jerarquía de interacciones, empezando por la formación de subunidades individuales y autónomas que se asocian entre sí por interacciones de difusión-colisión con otras subunidades, conformando posteriormente el cuarto o la mitad del barril alrededor del núcleo hidrofóbico. La propuesta fue apoyada por el descubrimiento de que mitades de barril pueden intercambiarse casi indiscriminadamente sin comprometer el plegamiento (Höcker 2004). Las enzimas HisF y HisA de *Thermotoga marítima*, que tienen plegamiento de barril TIM y son casi simétricas en su extremo carboxilo y amino, fueron fragmentadas para realizar las quimeras de la mezcla o duplicación de las mitades. En ellas se encontró una variante con mayor estabilidad que las mitades originales (Höcker 2004), con lo cual se reiteró la idea de que los barriles TIM pudieron evolucionar de antecesores compuestos de  $(\beta/\alpha)_4$ , como se comentó en la introducción.

En el 2006, el grupo que complementó el andamiaje de la CspA incursionó en la formación de proteínas a partir de polipéptidos sin la capacidad de plegarse por sí mismos, experimento que se comentó en la introducción (Riechmann 2006). El trabajo constó en complementar un fragmento de una lisozima del huevo de gallina con interfase de unión a unos anticuerpos específicos. El fragmento por sí solo, al no plegarse, era rápidamente digerido por proteólisis y, por otro lado, se unía débilmente con los anticuerpos. Al ligar aleatoriamente el polipéptido con fragmentos genómicos y seleccionando quimeras estables y funcionales, se encontró un dímero formado por el polipéptido unido a un grupo hemo con estabilidad hasta los 59 °C y aún mayor afinidad a los anticuerpos que el polipéptido aislado. Con estos resultados se comprobó que la unión de fragmentos de subdominios pudo ser la vía de evolución a los dominios (Riechmann 2006). También en el 2006, Boubarkri *et al.*, (Boubakri 2006) utilizaron la recombinación homóloga en el sentido inverso a la mayoría de los experimentos de la ingeniería de proteínas. En vez de intentar generar una nueva proteína, su objetivo fue reconstituir un gen, *linA*, el cual degrada lindano –un pesticida– y para el cual no se han encontrado homólogos. La reconstrucción se llevó a cabo a partir del ensamblaje de bloques con similitud a regiones del gen provenientes de DNA metagenómico obtenido de tierras sin lindano, es decir, de una población carente del gen *linA*. Tras la amplificación y el barajeo de los bloques se logró reconstituir hasta el 73% de *linA*, y, a partir de un banco de 1300 clonas complementadas con fragmentos del gen (obtenido de una cepa que sí lo presentaba), lograron generar 23 variantes activas. Con este trabajo, además de constatar que los genomas hay, si no infinita, si una gran fuente de innovación proteica, probaron que la recombinación conjunto a la transferencia lateral pueden generar variantes funcionales ante una presión ambiental. El mismo año, en nuestro grupo se generó la complementación de la PRAI<sub>Ec</sub> carente del fragmento  $(\beta\alpha)_3\beta_4$  –implicado en la nucleación temprana– con una poza de fragmentos del genoma de *E. coli* (Sánchez 2006). De  $10^3$  recombinaciones, el 2% fueron quimeras plegadas (Sánchez 2006). Con este trabajo se reitera que la conformación por modularidad hace de los plegamientos estructuras muy flexibles que potencian la versatilidad proteica. En conclusión, la evolución dirigida por barajeo de módulos es una oportunidad para generar librerías que representan líneas de microevolución en las cual se puede generar la diversidad que permite estudiar la función, el plegamiento e incluso la historia evolutiva.

## 4. OBJETIVOS

### Objetivo General.

Generar y evaluar una metodología para crear un banco de enzimas PRAI<sub>Ec</sub> (fosforibosil antranilato isomomerasa) quiméricas por su complementación a partir del barajeo de subdominios de homólogos distantes obtenidos a partir de la amplificación de una muestra ambiental de DNA.

### Objetivos particulares.

1. Complementar la PRAI de *E. coli* con fragmentos de tamaño de subdominios de secuencias amplificadas por oligonucleótidos dirigidos a genes parálogos de la PRAI.
2. Diseñar, a partir de análisis de áreas de secuencia conservadas en superfamilias de barriles TIM, algunos oligonucleótidos degenerados para la amplificación de genes ortólogos de la PRAI de *E. coli*. Probarlos en *E. coli* K12.
2. Analizar la capacidad y las constricciones metodológicas de generar diversidad por este método.

### 4.1 HIPÓTESIS.

Se espera que el sistema de complementación de la PRAI<sub>Ec</sub> por ligación de fragmentos obtenidos por PCR permita la generación de bancos de mutantes con capacidad de plegarse. Por otra parte, se espera amplificar barriles TIM de *E. coli* K12 utilizando los oligonucleótidos degenerados diseñados. Debido a su degeneración y a que el diseño de los mismos no se dirige a *E. coli* en particular, no se espera que todos los amplificados correspondan a barriles TIM ni que con todos los pares de oligonucleótidos probados se amplifiquen miembros de las superfamilias con las cuales se basó su diseño.

### 4.2 JUSTIFICACIÓN.

El presente trabajo tiene la virtud de abordar interrogantes evolutivas e innovaciones en la ingeniería de proteínas bajo un modelo prometedor, los barriles TIM. Por otra parte, el sistema propuesto busca generar diversidad sin comprometer la estabilidad de las proteínas en cuestión y eliminar la restricción en la variedad de resultados finales al simular la recombinación no homóloga a partir de múltiples pozas génicas. La utilización de homólogos distantes busca aumentar la probabilidad de acumulación de mutaciones sin introducir efectos deletéreos sin comprometer la exploración del espacio de secuencia ni del paisaje de adecuación debido a la baja similitud presente entre las superfamilias de barril TIM a nivel de secuencia, de estructura y de función. Por último, la metodología que se plantea busca aprovechar la diversidad de secuencias de

DNA y genomas proveniente de organismos no secuenciados y/o no cultivables a partir de la inclusión de DNA metagenómico.

## 5. MATERIAL Y MÉTODOS

### 5.1 DISEÑO DE OLIGONUCLEÓTIDOS Y PCR.

*Diseño de oligonucleótidos.* El diseño de **oligonucleótidos específicos** para amplificar parálogos de la PRAI<sub>Ec</sub> se realizó a mano a partir de secuencias representantes de cuatro superfamilias de barriles TIM (subtraídas del CATH, topología 3.20.20) a las cuales se les buscó su ortólogo en *E. coli* en el servidor del National Center for Biotechnology Information (NCBI) por Basic Local Alignment Search Tool (Blast, en este caso Blastp; Altschul 1997) (corte de valor de expectancia *v.e.*  $\leq 10^{-3}$ ). Los **oligonucleótidos degenerados** se diseñaron al obtener las secuencias de los dominios representantes de las 21 superfamilias de CATH v3.0.0 de los barriles TIM y buscar por Blastp (*v.e.*  $\leq 10^{-3}$ ) los ortólogos en *E. coli*. Los ortólogos de las secuencias de *E. coli* en los genomas bacterianos disponibles en el NCBI (2005) se detectaron con el criterio del mejor resultado recíproco (*bidirectional best hit*, el cual se describe brevemente en la sección 6.2) usando Blastp (*v.e.*  $\leq 10^{-3}$ ). Se realizaron alineamientos múltiples de los ortólogos de cada superfamilia en Muscle v3.41 (Edgar 2004) con las secuencias proteicas obtenidas y se caracterizaron por ALISTAT (Eddy 1998). Los oligonucleótidos se diseñaron por comparar los resultados de CODEHOP (Consensus-Degenerate-Hybrid Oligonucleotide Primers; Rose 2002), y de MEME 3.5.0 (Timothy 1998). CODEHOP es un hipervínculo del programa BMAP en el cual se buscan bloques de conservación y a partir de los cuales se proponen oligonucleótidos degenerados para cada bloque. Se utilizó el código genético estándar, la tabla de utilización de codones equivalente (*equal*) y los parámetros según los valores establecidos de inicio con excepción de la degeneración (la cual se varió de 128 a 256) y de la precisión (*strictness*; utilizada en un rango de 0.0 a 1.00). En MEME se buscan motivos y se presentan las matrices de probabilidad posición-específicas de las áreas conservadas, con las cuales se pueden diseñar los oligonucleótidos degenerados. Los parámetros para la búsqueda de motivos en MEME fueron que éstos ocurrieran de 0-1 veces por secuencia, que tuvieran de 25-50 aminoácidos y que encontrara un máximo 10 motivos por alineamiento. Todos los oligonucleótidos (oligos) diseñados para este trabajo fueron optimizados para los criterios de temperatura de hibridización (TM), longitud, porcentaje de guaninas y citosinas (%GC) y propensidad a dimerizarse ( $\Delta G$ ), analizados en el servidor Oligo Analyzer (Integrated DNA technologies 2006; <http://www.idtdna.com/analyzer/Applications/OligoAnalyze>), así como por la representatividad de la secuencia amplificada (porcentaje de nucleótidos amplificados del gen blanco amplificado) y el valor de expectancia de un Blastn de la secuencia del par de oligos (5'-oligo-3'NN5'-oligo\_reverso\_complementario-3') contra el gen blanco en *E. coli* K12 para los oligos específicos y contra las bacterias para los degenerados. Los oligos fueron sintetizados en la Unidad de Síntesis

del Instituto de Biotecnología, UNAM. La lista de oligos sintetizados y utilizados a lo largo del trabajo se encuentra en el Apéndice 1.

*Reacción en cadena de la polimerasa.* Las PCRs (termociclador iCycler BIO-RAD) inicialmente se realizaron con un gradiente de temperatura de extensión de 40-60 °C para determinar la  $T_M$  óptima de cada par de oligos, la cual, una vez determinada, se fijó para reacciones subsecuentes. Por lo demás, las reacciones se llevaron a cabo con 5 unidades (u) de Taq-DNA polimerasa (*Roche*), 40 pmol de cada oligonucleótido, Buffer Taq 10x (*Roche*) y 1.5 mM dNTPs (*Roche*) por cada 100  $\mu$ l volumen final de la reacción. El programa utilizado en todas las reacciones constó de un ciclo de desnaturalización a 94° 5 min, 25 ciclos de amplificación de 94 °C 50 seg, la  $T_M$  óptima 50 seg y 72 °C 1 min, finalizando con la extensión final a 72 °C por 5 min. Los productos de PCR se purificaron utilizando el kit de *High Pure PCR Product Purification Kit (Roche)* en solución. En algunos casos los segmentos amplificados fueron clonados con el kit *pGEM-T Vector System (Promega)*, y transformados en la cepa *E. coli* JM101 para confirmar la identidad del producto de PCR por secuenciación.

## 5.2 SISTEMA DE COMPLEMENTACIÓN DE LA PRAI<sub>Ec</sub> CON PARÁLOGOS.

La metodología constó en obtener, por un lado, algunos genes parálogos de la PRAI<sub>Ec</sub> de distintas superfamilias de Barril TIM, fragmentarlos y generar extremos rasos (reparar) y, por el otro lado, preparar el plásmido al digerir el gen de la PRAI<sub>Ec</sub> y de igual manera generar extremos rasos para así poder asociar covalentemente el inserto y el vector. Consecutivamente se tamizaron las variantes para seleccionar vectores con los genes de las PRAI<sub>Ec</sub> quiméricas con capacidad de plegarse por medio de un reportero de plegamiento a base del gen de la proteína Carboxil Acetil Transferasa (CAT) río abajo del gen de la PRAI<sub>Ec</sub>. A continuación se describe el sistema del tamizaje por el reportero de plegamiento y los materiales y métodos divididos en sus etapas principales: la preparación de los fragmentos, del plásmido, la ligación y transformación, el tamizaje, y el análisis de las secuencias por métodos bioinformáticos, por SDS-PAGE y Western Blot.

*Tamizaje usando el reportero de plegamiento CAT.* El sistema de tamizaje de proteínas plegadas se basa en la fusión del gen cloranfenicol acetil transferasa (*cat*) al gen a modificarse y consecuentemente a la fusión de las proteínas resultantes de la traducción de ambos genes. El gen *cat* codifica para una enzima homotrimérica de 25 kDa que inactiva al cloranfenicol, antibiótico que se une a los ribosomas 50s, bloqueando la peptidiltransferasa. Cuando la CAT está fusionada a una proteína insoluble, exhibe un decremento en la resistencia a cloranfenicol; mientras más resistencia

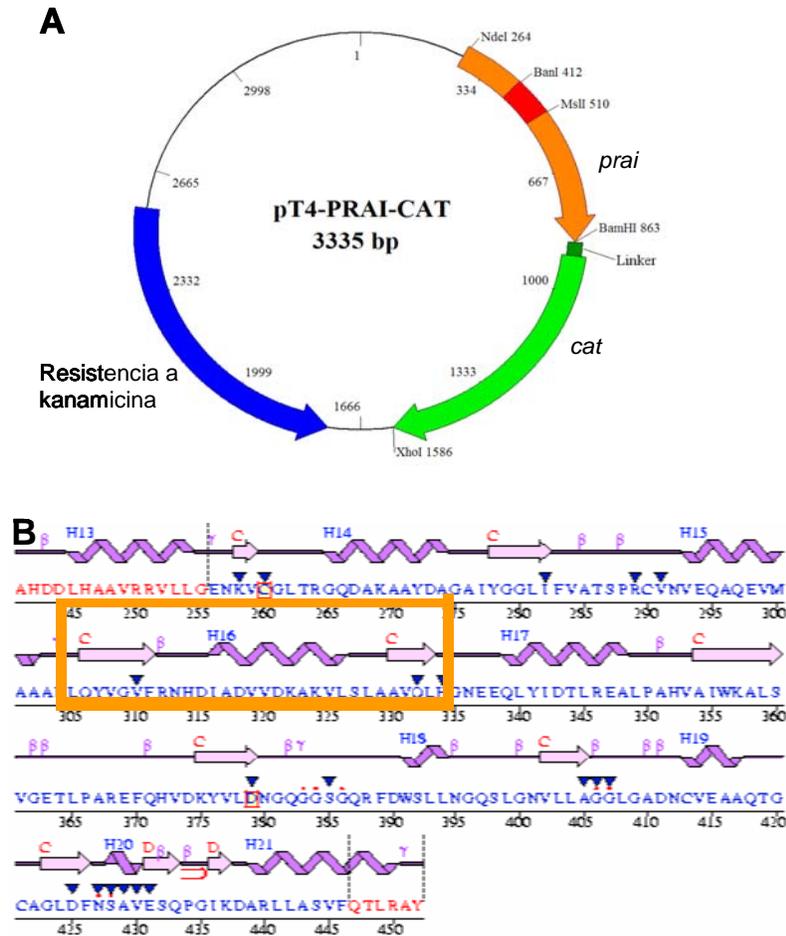
presente, la fusión es más estable. La solubilidad *in vivo* de las proteínas mutantes, por tanto, puede ser determinada por la resistencia de las cepas al antibiótico y *Western blots* para comprobar la fusión entre las proteínas y evitar falsos positivos (Maxwell 1999).

*Fragmentación y reparación de los insertos.* Los productos obtenidos de las PCRs por oligos específicos fueron fragmentados por sonicación (Branson Sonifier 450 prestado por X. Soberón, ~50 µg de muestra en 400 µl de volumen final (Vf), output control 3, ciclo constante, de 1 a 10 aplicaciones de 10 seg, manteniendo 3 min en hielo durante las aplicaciones) o por DNaseI (*Roche*, libre de RNase; 80 µg de muestra, 10 µl de Buffer PCR –MgSO<sub>4</sub> 10x, MnCl<sub>2</sub> 1 mM pH =7.4, 1.5 µl de DNase (5 µg/ml en PBS), 100 µl Vf, cinética a 16 °C, detener con 5 µl EDTA 100 mM). Aquellos fragmentos entre 100 y 300 pb fueron purificados por gel de agarosa al 2% con el kit *Agarose Gel DNA Extraction Kit (Roche)*. La reparación de los fragmentos constó en modificar los extremos en cadena sencilla producidos por la sonicación (la DNase con MnCl<sub>2</sub> dejan extremos rasos) en extremos rasos, lo cual se realizó con las polimerasas Klenow y T4 DNA polimerasa (T4 pol) rellenando los extremos 3' recesivos y protuberantes, respectivamente (Apéndice 2).

*Preparación del plásmido.* El vector, construido en el laboratorio del Dr. Hilvert, ETH, Suiza y donado por el Dr. Joel Osuna al laboratorio donde posteriormente se modificó para trabajos previos relacionados, es el plásmido pT4-PRAI-CAT (3360 pares de bases, pb) con resistencia a kanamicina (*Km*), origen de replicación pUC y un promotor transcripcional medianamente fuerte *ptrc*. Éste contiene al gen de la PRAI<sub>Ec</sub> en su estado monomérico seguido por un linker y por el gen de la CAT, por lo cual resiste a cloranfenicol (*Cm*) en su fenotipo silvestre o *wild tipe (WT)* (Figura 5). El sitio de digestión de la PRAI, ( $\beta\alpha$ )<sub>3</sub> $\beta$ <sub>4</sub>, fue seleccionado en un trabajo previo (Sánchez 2006) basándose en el programa SCHEMA (Voigt 2002) y un alineamiento múltiple de 12 homólogos de la PRAI (Sánchez 2006). Con el primer análisis se seleccionaron los sitios con menor probabilidad de comprometer el plegamiento al ser eliminados; con el alineamiento múltiple se restringió la selección al sitio que está más conservado (Sánchez 2006). La región seleccionada comprende del residuo 51 al 83 (*ie.* 93 pb), con tres aminoácidos del sitio activo. Se ha sugerido que el fragmento es un sitio de nucleación importante para los eventos tempranos del plegamiento (ver sección 2.3.1) (Akanuma 2005).

Antes de cada ciclo de preparación, el vector pT4-PRAI-CAT, transformado en la cepa XL1blue-MRF' se replaqueó, se creció por 16 hrs (*ON*) a 37 °C en 3 ml de medio líquido Luria Bertani (LB; *Km* 35 µg/ml) y se purificó por lisis alcalina con el kit *High Pure Plasmid Isolation Kit, Roche*. Se realizaron digestiones dobles con las enzimas BanI y MslI *ON* a 37 °C (Apéndice 3).

La reparación del plásmido constó en cambiar los extremos pegajosos obtenidos por la enzima *BanI* por extremos rasos, lo cual se realizó del mismo modo que la reparación de fragmentos o con la enzima exonucleasa Mung-Bean (Apéndice 2). Se desfosfataron los extremos del plásmido con el fin de disminuir el número de vectores autoligados con la enzima CIP (Calf Intestinal Phosphatase) (Apéndice 4). La purificación del vector entre reacciones así como al final de su preparación se realizó con el kit *High Pure PCR Product Purification, Roche* o por fenol-cloroformo (Sambrook 1989).



**Figura 5. A)** Representación circular del plásmido PT4-PRAI-CAT utilizado en este trabajo (Sánchez 2006). **B)** Secuencia de aminoácidos y su conformación espacial en estructuras secundarias de la fosforibosil antranilato isomerasa de *E. coli* (sin presentar completa la secuencia previa correspondiente a la enzima IGPS, con la cual se encuentra fusionada de manera natural). Se muestran los sitios activos (triángulos azules), los residuos catalíticos (aminoácidos en recuadros rojo con amarillo) y los sitios de interacción con el ligando (puntos rojos). El fragmento seleccionado con el recuadro naranja es el sitio a complementar (Modificado de PDB).

*Ligación y transformación.* Para las ligaciones se utilizó la T4-DNAligasa (*Roche*) ON a 16 °C. La concentración final y la relación molar entre el inserto y el vector varió en cada experimento. La

reacción se purificó previamente a su transformación por extracción con butanol y precipitación con etanol (Apéndice 5), resuspendiendo en 10  $\mu$ l de agua tetradestilada. La transformación al electroporar (1.8 V en electroporador Bio-Rad ajustado a 25  $\mu$ F y 200  $\Omega$ ) alrededor de 100 ng de la ligación en cepas de XL1Blue-MRF o MC1061 y recuperando en 1 ml de SOC (Sambrook 1989). Se utilizó pUC 18 Gilbert W, 1991 para probar la eficiencia de la transformación plaqueando 50  $\mu$ l en cajas con LB y Ampicilina (*Am*) a 200  $\mu$ g/ml

*Tamizaje.* El tamizaje comenzó seleccionando las células transformadas en placas con medio de cultivo Luria Bertani (LB) sólido y *Km* 25  $\mu$ g/ml *ON* a 37 °C por ser la resistencia del plásmido en placas con LB y *Cm* a 15  $\mu$ g/ml *ON* a 30° para identificar quimeras con proteínas plegadas. Las células resistentes a *Cm* fueron crecidas en 3 ml de LB líquido con *Km* 25  $\mu$ g/ml por 8 hrs a 37 °C tras lo cual se extrajo su DNA por choque térmico (Sambrook 1989) para realizar PCRs (con el programa descrito anteriormente) con oligos NdeI y BamHI (Apéndice 1) fuera del sitio deletado para comparar los tamaños de los insertos. Se extrajo el plásmido por lisis alcalina con el kit *High Pure Plasmid Isolation Kit, Roche* de aquellas clonas prometedoras y se mandaron a secuenciar con los oligos PT4 y CAT (Apéndice 1) en la Unidad de Síntesis y Secuenciación Instituto de Biotecnología, UNAM.

*Identificación del inserto clonado y predicción de estructuras secundarias.* Las secuencias obtenidas se alinearon en DNA Strider 1.4f6 para identificar los fragmentos insertados, a los cual se les hizo un Blast contra el genoma de *E. coli* K12 para identificar el gen de origen. Se predijeron las estructuras secundarias de las quimeras completas, del inserto en su fase original y del inserto independiente en el programa PSIPRED v2.5 (Bryson 2005).

*Concentración mínima inhibitoria (CMI).* Con el experimento de la CMI se calcula la concentración de un antibiótico a la cual la mitad de las clonas analizadas no son resistentes, por lo que se considera como la concentración del antibiótico a la que una cepa dada resiste. Las clonas resistentes a *Cm* (*Cm*<sup>R</sup>) y en fase con la PRAI<sub>Ec</sub> fueron retransformadas en la cepa JM101; como control positivo se utilizó el pT4-PRAI-CAT (CMI ~140  $\mu$ g/ml) transformado en la misma cepa; como control negativo la cepa sin plásmido (CMI ~5  $\mu$ g/ml). Se realizaron diluciones del cultivo recuperado de 10<sup>0</sup> a 10<sup>-8</sup> con LB. Se sembraron 5  $\mu$ l de las diluciones de cada cepa en cajas con *CM* en concentraciones de 0, 5, 10, 15 y 45  $\mu$ g/ml, así como en una caja de *Km* 25  $\mu$ g/ml. Las cajas se incubaron *ON* a 30 °C.

*Gel de proteínas SDS-PAGE.* Se hicieron cultivos *ON* a 30 °C y 250 rpm de 3 ml LB con *CM* 15 µg/ml de las clonas seleccionadas, el pT4-PRAI-CAT y de la cepa JM101 sin antibiótico. Se centrifugó 1 ml de cada ensayo 5 min a 4000 rpm, se eliminó el sobrenadante y se resuspendió el pellet en 500 µl de PBS. Las muestras se sonicaron (Branson Sonifier 450) 10 seg 2 veces cada una (control output 6), agregando 1 µl de PMSF 100 mM tras la primera sonicación. Se centrifugaron a 4000 rpm 10 min, recuperando el sobrenadante (fracción soluble) y resuspendiendo el pellet en 100 ml de mezcla de digestión (fracción insoluble). Las muestras fueron observadas en un gel de acrilamida al 12.5%.

*Western Blot.* Las proteínas del gel SDS-PAGE de la fracción soluble realizado como se describió anteriormente se transfirieron a una membrana de nitrocelulosa. La membrana fue teñida con Ponceau para comprobar la transferencia y posteriormente bloqueada *ON* a 4 °C con solución TBST-leche (Svelty, Nestlé) 5%. Después de lavar tres veces con TBST (10 min cada vez) se incubó la membrana por 1 hora con el anticuerpo ANTI-CAT (*Roche*) en una relación 1:3000. Se lavó tres veces más con TBST y leche al 0.1% y se incubó 1 hora con el anticuerpo ANTI-DIG (*Roche*) en relación 1:5000. Tras dos lavadas con TBST se adicionaron 15 ml de H<sub>2</sub>O con 1 ml BCIP/NTP (*Zymed*) 1:1 hasta visualizar las bandas y se suspendió la reacción con EDTA 0.005 M.

### 5.3 OTRAS CONSIDERACIONES METODOLÓGICAS.

Para visualizar productos o purificar por gel se utilizaron geles de agarosa grado DNA (BIO-RAD®) al 0.8, 1 o 2% (en relación inversa al tamaño de la secuencia en cuestión) en solución amortiguadora TAE 0.8 X (Sambrook 1989).

Las células electrocompetentes XL1Blue-MRF y JM101 (calcio competentes) se prepararon según Sambrook 1989.

Se presenta a continuación el fenotipo de las cepas utilizadas:

XL1-Blue: *F'::Tn10 proA+B+ lacIq Δ(lacZ)M15/recA1 endA1 gyrA96 (Nal<sup>r</sup>) thi hsdR17 (rK-mK+) glnV44 relA1 lac.*

JM101 Ca: *F' traD36 proA<sup>+</sup>B<sup>+</sup> lacIq Δ(lacZ)M15/ Δ(lac-proAB) glnV thi.*

## 6. RESULTADOS Y DISCUSIÓN

### 6.1 COMPLEMENTACIÓN DE LA PRAI<sub>Ec</sub> POR FRAGMENTOS DE GENES PARÁLOGOS.

Se eligieron cuatro proteínas con dominios parálogos a la PRAI<sub>Ec</sub> de distintas superfamilias de barriles TIM para probar el sistema de complementación al azar con fragmentos de genes amplificados por PCR y tamizaje por fusión a CAT. Las cuatro proteínas fueron las siguientes: una alanina racemasa (*alr*), una xilosa isomerasa (*xyl*), una tRNA-guanina transglicosidasa (*tRg*) y una homocisteína metiltransferasa (*hmt*). La longitud e identificadores de sus genes, así como el número de superfamilias (CATH v3.0.0) a las cuales pertenecen se presentan en la Tabla 1. Para cada una de estas proteínas se diseñaron un par de oligonucleótidos (oligos) de manera que se amplificara la mayor parte de la secuencia de sus genes. La secuencia de los oligos sintetizados se muestran en el Apéndice 1; sus propiedades se presentan en la Tabla 2.

**Tabla 1.** Información de las proteínas parálogas utilizadas para complementar a la PRAI.

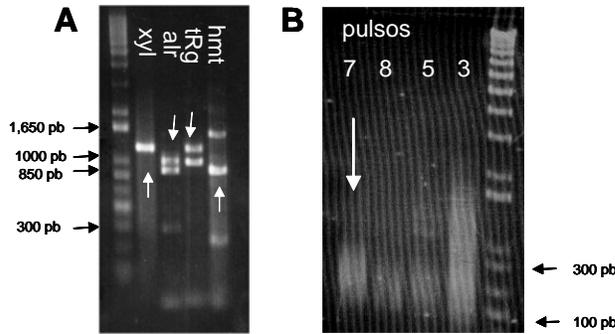
| Clave      | Gen ( <i>pb</i> ) | ID                      | Superfamilia |
|------------|-------------------|-------------------------|--------------|
| <i>alr</i> | 1077              | NP_418477.1 GI:1613187  | 3.20.20.10   |
| <i>xyl</i> | 1320              | NP_418022.1 GI:16131436 | 3.20.20.150  |
| <i>tRg</i> | 1125              | NP_414940.1 GI:16128391 | 3.20.20.105  |
| <i>hmt</i> | 930               | NP_414795.1 GI:16128246 | 3.20.20330   |

Gen (*pb*), longitud en pares de bases del gen de la proteína en cuestión. Superfamilia, clave de superfamilia homóloga de las proteínas (CATH v3.0.0).

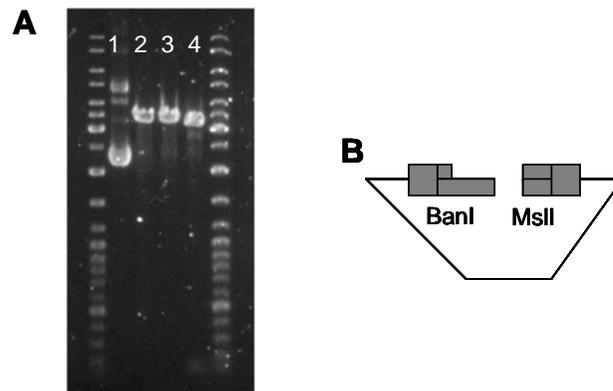
**Tabla 2.** Características de los oligos diseñados para la amplificación específica de parálogos de la PRAI<sub>Ec</sub>.

| Oligo            | Tm práctica | longitud oligo ( <i>pb</i> ) | %GC  | ΔG (kcal/mol) | <i>pb</i> amp |
|------------------|-------------|------------------------------|------|---------------|---------------|
| <i>alr</i> 5'-3' |             | 18                           | 50   | 3.61          |               |
| <i>alr</i> 3'-5' | 56.2        | 22                           | 45.5 | 6.3/5.47      | 1080          |
| <i>xyl</i> 5'-3' |             | 22                           | 50   | 3.61          |               |
| <i>xyl</i> 3'-5' | 62          | 24                           | 58.3 | 5.02/6.69     | 1224          |
| <i>tRg</i> 5'-3' |             | 19                           | 52.6 | 5.02          |               |
| <i>tRg</i> 3'-5' | 52.6        | 24                           | 54.2 | 6.75/5.47     | 1034          |
| <i>hmt</i> 5'-3' |             | 20                           | 55   | 7.06          |               |
| <i>hmt</i> 3'-5' | 60          | 21                           | 66.7 | 7.06/9.76     | 868           |

Tm, temperatura de hibridación; *pb*, pares de bases; %GC, porcentaje de guaninas y citocinas en la secuencia; ΔG, ΔG homodimerización/ΔG heterodimerización; *pb* amp, *pb* a amplificar por PCR.



**Figura 6.** Preparación de los insertos: amplificación de genes parálogos de la PRAI<sub>Ec</sub> en *E. coli* y cinética de sonicación. **A)** Resultados de PCRs de oligos específicos para amplificar en *E. coli* parálogos de la PRAI. Los productos blanco de los pares de oligos (señalados con una flecha) xyl, alr, tRg y hmt son en *pb* de 1224, 1080, 1034 y 868, respectivamente **B)** Cinética de sonicación (ciclos constantes, output control 3) con 3, 5, 7 y 8 pulsos. La muestra sonicada por 7 pulsos fue la que más productos entre 100 y 300 *pb* generó, por lo cual se eligió para posteriores sonicaciones.



**Figura 7.** Preparación del vector: Digestión. **A)** De izquierda a derecha, 1) vector sin digerir (pT4-P-C), 2) digestión sencilla con BanI (la cual deja extremos pegajosos), 3) digestión sencilla con MspI (la cual deja extremos rasos) y 4) digestión doble (extremo pegajoso - extremo raso). **B)** Representación del vector digerido con un extremo 3' recesivo y un extremo raso.

Se estandarizaron las PCRs por gradientes de temperatura de hibridación. Para los cuatro pares de oligos se encontró la banda del tamaño esperado correspondiente al fragmento del gen parálogo blanco conjunto a dos o tres bandas más, productos inespecíficos de la reacción (Figura 6A). Se realizó una cinética de sonicación para el conjunto de los fragmentos amplificados por los cuatro pares de oligos para establecer el protocolo en el cual se maximizaran los productos entre los 100 y 300 *pb* –aunque los productos de la sonicación de estos tamaños fueron purificados posteriormente por gel de agarosa para minimizar la ocurrencia de fragmentos de otros tamaños– (Figura 6B). Después de la sonicación, los fragmentos se repararon (*ie.* los extremos protuberantes se eliminaron) con las enzimas Klenow y T4 pol, generando así los insertos. No se excluyeron de los experimentos de complementación los fragmentos inespecíficos amplificados en la PCR debido a que el objetivo de este trabajo es *aumentar la probabilidad* del éxito de recombinaciones no

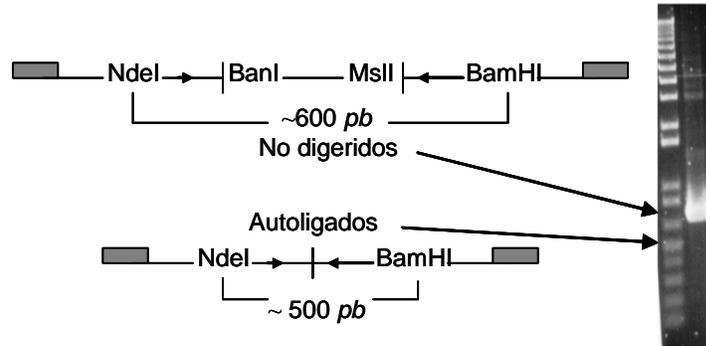
homólogas para generar variantes proteicas basándose, pero no restringiéndose, en el barajeo de homólogos distantes.

Concomitantemente, el vector fue digerido con las enzimas BanI y MslI para eliminar el fragmento  $(\beta\alpha)_3\beta_4$  (ver sección 5.2; Figura 5B) y sus extremos fueron reparados y desfosfatados como se describe en la metodología (Figura 7). Posteriormente a la ligación del plásmido con los insertos y a su transformación en células competentes, se tamizó para obtener cepas con resistencia a kanamicina ( $Km^R$ ; resistencia del plásmido) y cepas con resistencia a kanamicina y a cloranfenicol ( $Cm^R$ ), cuyo fenotipo es el de una fusión de la  $PRAI_{EC}$  (quimérica o *wt*) con la CAT. Las cepas  $Km^R$  -  $Cm^R$  fueron analizadas por PCR para determinar por el tamaño del amplificado si se trataban de falsos positivos o quimeras a caracterizar. A continuación se presentan los resultados divididos en los dos primeros bancos generados, controles de la preparación del plásmido, un banco más y la caracterización de las clonas quiméricas.

*Resultados del Banco A: obtención mayoritaria de plásmido no digerido.* En un primer intento para generar el banco de  $PRAI_{EC}$  quiméricas se realizaron dos ligaciones: una ligación en relación molar 1:3 (60 ng totales) de vector-inserto (lig/*v-i*) y, como control negativo, una autoligación de vector ( $_{ALV}$ ). Las ligaciones se transformaron en células con eficiencia de  $1 \times 10^8$  transformantes/ $\mu$ g DNA, con lo cual se obtuvo un banco de 3288 clonas (total de cepas lig/*v-i*  $Km^R$  – total de cepas en  $_{ALV}$   $Km^R$ ). De las 30 clonas  $Cm$  resistentes ( $Cm^R$ ) analizadas por PCR, todas fueron clonas con vector no digerido (Figura 8). Para mejorar la eficiencia de obtención del vector digerido se realizó una cinética de digestión. A partir de las 2 hrs de incubación no hubo relación observada en geles de agarosa entre el tiempo de incubación y la cantidad de vector digerido.

*Resultados del banco B: obtención mayoritaria de plásmido autoligado.* Para el segundo banco se decidió seguir incubando la digestión *ON* y, para optimizar la obtención de vectores no superenrollados, las digestiones se purificaron por geles de agarosa corridos el tiempo suficiente para posibilitar la distinción entre los distintos estados del vector. Por otra parte, el vector se reparó y desfosfató de la misma manera. Como resultado de la transformación de la ligación vector-inserto en relación molar 1:3, (en células con eficiencia de  $1 \times 10^7$  transformantes/ $\mu$ g DNA) se obtuvieron 80 colonias  $Km^R$  de  $_{ALV}$  y 100 colonias  $Km^R$  de lig/*v-i*. No se obtuvieron cepas  $Cm^R$ . De las 25 colonias  $Km^R$  revisadas (provenientes de lig/*v-i*), 10 fueron colonias con vectores no digeridos y 15 colonias con vectores digeridos y religados. La eficiencia de obtención de vector digerido aumentó –por encontrar mayor porcentaje de clonas con vectores autoligados que no digeridos–, pero no se maximizó. Por otra parte, la presencia de algunas colonias autoligadas denotó un problema en la

desfosfatación de los extremos del plásmido y el nulo rendimiento de colonias con inserto tanto en las placas con *Km* como con *Cm* sugirió una baja eficiencia en la reparación del plásmido, en la ligación y/o baja calidad de los insertos.

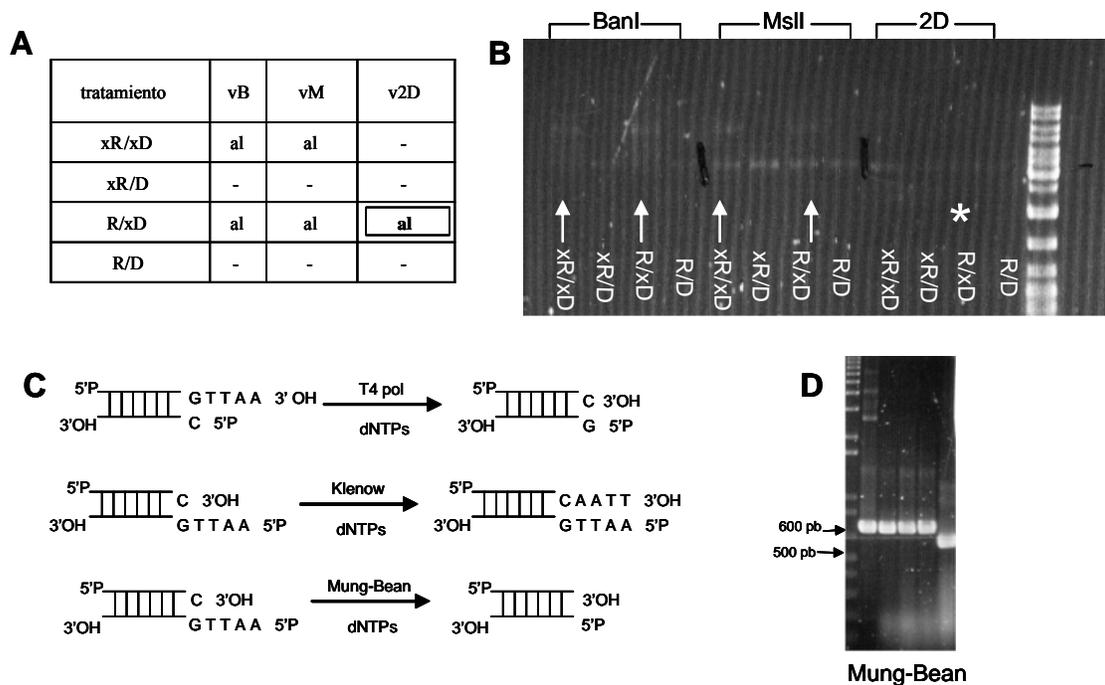


**Figura 8.** Tamizaje de clones *Km<sup>R</sup>* o *Cm<sup>R</sup>* por PCR. Los vectores purificados de las clonas de interés son amplificados con los oligos NdeI y BamHI para determinar si corresponden a una complementación, autoligación o no digestión. El producto de amplificación del vector no digerido es de alrededor de 600 pb, el del vector autoligado de 500 pb y de un vector complementado de tamaño diferente a éstos. En el primer banco de las 30 clonas analizadas, todas presentaron vector no digerido (600 pb).

*Controles para detectar los problemas en la preparación del plásmido.* Para probar la eficiencia de la reparación y desfosfatación de los extremos del plásmido digerido se realizaron 12 autoligaciones cada una de 80 ng de vector tratado distintamente. Los vectores fueron previamente digeridos con BanI (*vB*; enzima que deja extremos pegajosos), con MspI (*vM*; enzima que deja extremos rasos) o con ambas enzimas a la vez (*v2D*) y autoligados sin reparar ni desfosfatar (*xR,xD*), solamente reparados (*R,xD*), solamente desfosfatados (*xR,D*) o reparados y desfosfatados (*R,D*) (Figura 9A). La propensidad de cada vector a autoligarse se evaluó en geles de agarosa por la presencia de bandas de mayor tamaño a la del vector digerido. El único resultado que se contrapuso con lo esperado (la propensidad de autoligarse teóricamente correcta) fue el de la autoligación del *v2D* reparado y no desfosfatado (el cual se esperaba que se autoligara y no se autoligó), por lo cual se concluyó que la reparación del vector doblemente digerido no estaba funcionando satisfactoriamente (Figura 9B). Con estos controles se concluyó, por otro lado, que el problema de la obtención de colonias con vectores autoligados se debe a que no se desfosfatan los extremos del 100% de plásmidos tratados; el plásmido digerido con ambas enzimas y autoligado, sin embargo, no crece a la concentración de *Cm* con la que se seleccionó, por lo que la presencia de un residual de plásmidos con extremos no desfosfatados no representó un problema metodológico.

Se realizaron otros controles de reparación para *v2D* por protocolos con exonucleasa Mung-Bean (enzima que degrada DNA de cadena simple; Figura 9C) se presenta en el Apéndice 2. Las

autoligaciones fueron transformadas en células con eficiencia de  $1 \times 10^6$  transformantes/ $\mu\text{g}$  DNA y crecidas en *Km*. De los 5 vectores analizados tratados con exonucleasa Mung-Bean, cuatro de ellos fueron vectores no digeridos y uno de ellos fue autoligado y correctamente reparado (Figura 9D). Al mismo tiempo que se hicieron estos experimentos, en el laboratorio se logró complementar la  $\text{PRAI}_{Ec}$  por el mismo protocolo con el cual se inició el presente trabajo (doble digestión BanI-MsII, repreción por Klenow y T4 pol y desfosfatación con CIP). La única variante fue que en la reparación con Klenow y T4 pol en vez de utilizar 1.5  $\mu\text{l}$  de dNTPs al 2.5 mM se utilizaron 4  $\mu\text{l}$  (Sánchez 2006). Dado el éxito de la aplicación de este protocolo, se prosiguió utilizando Klenow y T4 pol (4  $\mu\text{l}$  de dNTPs 2.5 mM) en la generación de otros bancos.



**Figura 9.** Controles para determinar la eficiencia de la reparación y desfosfatación. **A)** Tabla de resultados esperados de las autoligaciones del vector digerido con BanI, MsII o por ambas enzimas (2D) y reparado (R,xD), desfosfatado (xR,D), reparado y desfosfatado (R,D) o sin tratamiento posterior a la digestión (xR,xD); *al*, tratamiento con el que se espera autoligación del vector; en recuadro y negritas el resultado que contrastó con lo esperado. **B)** Gel de agarosa de las ligaciones. Los tratamientos que produjeron autoligación del vector presentan una flecha blanca; la reparación de la doble digestión sin desfosfatar no permitió la autoligación del vector, al contrario de lo esperado (asterisco blanco). **C)** Actividad de las enzimas reparadoras. La T4 DNA polimerasa (T4 pol) remueve los extremos 3' protuberantes, Klenow rellena los extremos 3' recesivos (más eficientemente que T4 pol, la cual también tiene la misma actividad) y Mung-Bean remueve los extremos protuberantes 3' y 5'. **D)** una clona tratada con exonucleasa Mung-Bean resultó reparada satisfactoriamente. Las demás clonas analizadas presentaron vectores no digeridos.

*Resultados del banco C: fragmentación de insertos por DNasa y obtención de cuatro clonas quiméricas.* Para el siguiente banco los genes parálogos se fragmentaron por DNasa con  $\text{Mn}^{2+}$  en vez de por sonicación, de manera tal que no se requirieran reparar posteriormente y que se evitara el

daño físico que la sonicación puede ocasionar en los fragmentos de DNA. El vector fue preparado normalmente con el protocolo optimizado de reparación. Al contrario de los fragmentos sonicados, se observó por gel de agarosa autoligación de los fragmentos tratados con DNasa, lo que implicó que sus extremos estaban rasos (Figura 10). Después de transformar las ligaciones (relación 1:10 y 250 ng totales) en células con eficiencia de  $1 \times 10^8$  transformates/ $\mu\text{g}$  DNA se obtuvieron 20 colonias  $Cm^R$  provenientes de la lig/*v-i*. De los vectores analizados por PCR (purificados de las 20 cepas  $Cm^R$ ), 14 presentaron bandas del tamaño correspondiente al vector no digerido, 1 al tamaño de vector autoligado y 5 presentaron bandas de tamaño del rango esperado para las clonas quiméricas (Figura 11). Se secuenciaron 4 de estos vectores (dos presentaron el mismo tamaño de banda) de las cepas nombradas  $Cm^R$  1,  $Cm^R$  11,  $Cm^R$  14 y  $Cm^R$  15, a los cuales se les confirmó la presencia de inserto.

*Caracterización de las clonas quiméricas.* Los insertos de los 5 vectores están en fase con PRAI<sub>Ec</sub> y con CAT; en la Figura 12 se presentan las CMI de las clonas y en la Tabla 3 un resumen de sus características. La clona  $Cm^R$  1 tiene un inserto de 132 pb de un ensamble del oligo *hmt* reverso y presenta el nivel más bajo de CMI de las 5 clonas, resistiendo 15  $\mu\text{g}/\text{ml}$  de *Cm*. El inserto de la clona  $Cm^R$  11, la cual resistió 45  $\mu\text{g}/\text{ml}$  de *Cm*, es un fragmento del transportador de glutamato del sistema I GNT de gran afinidad (*gi* 49176350;  $v.e \leq 3 \times 10^{-22}$ ) –producto inespecífico de alguno de los pares de oligos– de 138 pb que, en su fase original forma 3  $\alpha$ -hélices. En la clona  $Cm^R$  14 se insertó un ensamble de 42 pb de los oligos del par para amplificar la tRNA guanina transglicosidasa; la clona resistió hasta 40  $\mu\text{g}/\text{ml}$  de *Cm*. Por último, la  $Cm^R$  15 presenta un inserto de 81 pb perteneciente al segmento  $\alpha/\beta/\alpha$  del dominio barril de hojas- $\beta$  de la alanina racemasa ( $v.e \leq 2 \times 10^{-38}$ ). En esta clona no se conservó el marco de lectura de la proteína de origen lo cual podría sugerir un desorden importante en las interacciones de aminoácidos y, sin embargo, presentó una CMI de 40  $\mu\text{g}/\text{ml}$  de *Cm* lo que indica un correcto plegamiento de la fusión PRAI<sub>EC</sub> $Cm^R$  15 – CAT.

En la predicción de la estructura secundaria de los cuatro insertos individuales se observa una presencia mayoritaria de  $\alpha$ -hélices y, en el contexto de las quimeras resultantes, una mayoría de  $\alpha/\beta$ ; sobresale en las predicciones de los insertos en contexto de las quimeras aquella de la clona  $Cm^R$  1, en la cual se predijo un fragmento  $\beta/\alpha/\alpha/\alpha$  (Tabla 3). Es posible que la baja resistencia a *Cm* de ésta clona esté relacionada con una complementación poco afín con la estabilidad y solubilidad del barril TIM.

Las 5 proteínas quiméricas se expresaron en *E. coli* JM101 (Figura 13); se confirmó su presencia en la fase soluble por un Western anti-CAT, con lo cual se ratificó también la unión del

reportero CAT a la quimera y por tanto, su correcto plegamiento y estabilidad en el citoplasma (Figura 14).

6.1.1 *Conclusiones y perspectivas.* El objetivo de los experimentos cuyos resultados se presentan en esta sección era complementar la  $PRAI_{EC}$  a partir de fragmentos de secuencias amplificadas por oligos dirigidos a parálogos de ésta proteína para probar el sistema (ver sección 4). Se lograron obtener 4 clonas quiméricas plegadas y estables en el citoplasma, por lo que el sistema se probó satisfactorio.

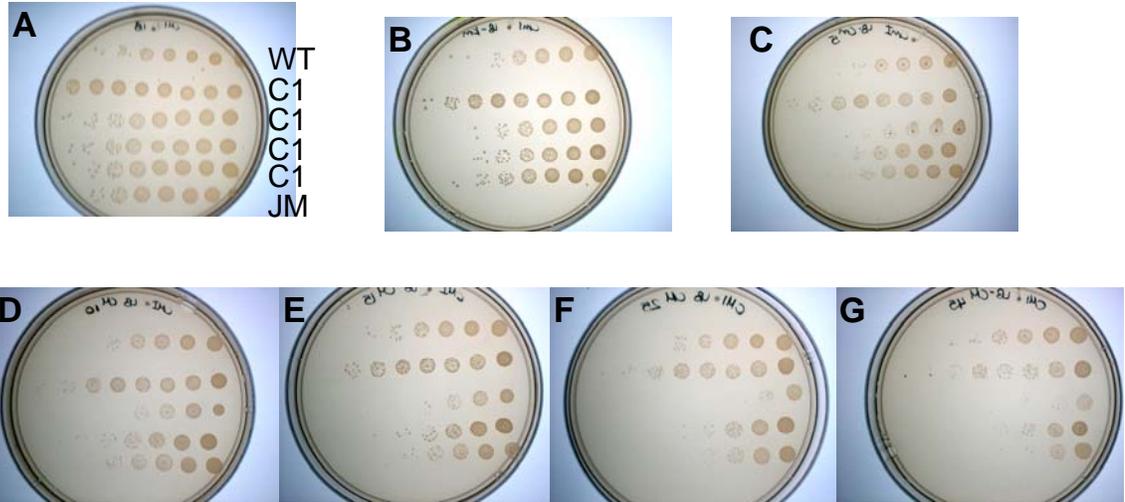
Para optimizar el método por mejorar la eficiencia de la doble digestión del plásmido, problema metodológico principal, se puede digerir el vector con una enzima a la vez, si, como en este caso, el producto de la doble digestión es del tamaño de una de las bandas del vector superenrollado. La generación de un cassette que inhabilite la reproducción del vector sin digerir puede sobreponer del todo el problema. Como se puede concluir en esta sección, el éxito del resto de la metodología depende de la optimización de las etapas individuales.

Un análisis más profundo de las clonas quiméricas obtenidas permitiría una conclusión más contundente. Es importante, por ejemplo, observar la solubilidad de las quimeras desfusionadas a CAT, o, más allá de probar el sistema, se podría determinar su estabilidad y presencia de estructuras secundarias y terciarias para entender la naturaleza (el tipo de plegamiento que adoptaron) y el por qué del éxito de las recombinantes generadas. Por otro lado, se podría dirigir la complementación únicamente a partir de fragmentos de genes parálogos para comparar los fenotipos de las quimeras generadas totalmente al azar con aquellas resultantes del barajeo a partir de homólogos distantes, y poder discernir cuantitativamente si éstas últimas, como aquí se plantea, aumentan las probabilidades del éxito de las recombinaciones no homólogas.

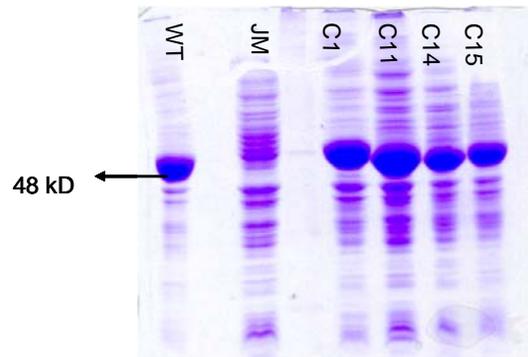
**Tabla 3.** Características de las quimeras obtenidas.

| Clona | Longitud inserto (aa) | Origen                  | CMI $C_m$ ( $\mu\text{g/ml}$ ) | ORF           | Psipred inserto              | Psipred origen         | Psipred quimera              |
|-------|-----------------------|-------------------------|--------------------------------|---------------|------------------------------|------------------------|------------------------------|
| B9C1  | 44                    | oligos 11t8             | 15                             | -             | $\alpha/\beta/\alpha/\alpha$ | -                      | $\beta/\alpha/\alpha/\alpha$ |
| B9C11 | 46                    | transportador gluconato | 45                             | conservado    | $\alpha/\alpha/\alpha$       | $\alpha/\alpha/\alpha$ | $\alpha/\beta/\alpha/\beta$  |
| B9C14 | 14                    | oligo 1r5y              | 40                             | -             | $\alpha$                     | -                      | $\alpha/\beta/\beta$         |
| B9C15 | 27                    | alanina racemasa        | 40                             | no conservado | $\alpha/\alpha$              | $\alpha/\beta/\alpha$  | $\beta/\alpha$               |

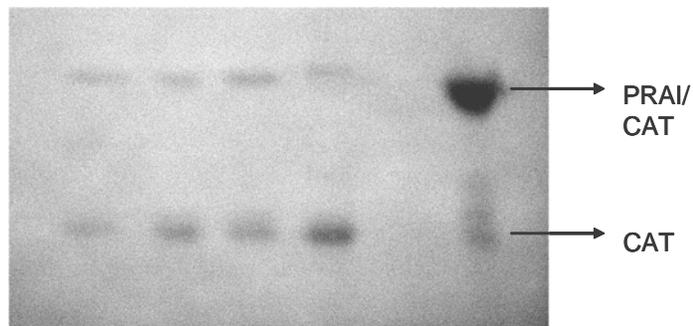
*CMI*, concentración mínima inhibitoria de las clonas a cloranfenicol ( $C_m$ ); *ORF*, marco de lectura abierta; *Psipred*, predicción de estructura secundaria por PSIPRED v2.5 (Bryson 2005) del inserto (Psipred inserto), del inserto en contexto del gen de origen (Psipred origen) y del inserto en contexto de la  $PRAI_{EC}$  complementada (Psipred quimera).



**Figura 12.** Concentración mínima inhibitoria (CMI) de las cuatro clonas quiméricas. De derecha a izquierda son diluciones de 100-108; de arriba abajo, cepa con vector de la *PRAIEc* WT, C11, C1, C14, C15 y la cepa *E. coli* JM101. **A)** LB sin antibiótico; **B)** LB/Km 25µg/ml; **C)** LB/Cm 5µg/ml; **D)** LB/Cm 10µg/ml; **E)** LB/Cm 15µg/ml; **F)** LB/Cm 25µg/ml; **G)** LB/Cm 45µg/ml.



**Figura 13.** Expresión de las quimeras *in vivo*. Gel de acrilamida de las proteínas de las fracciones insolubles de la cepa con el vector WT, la cepa *E. coli* JM101, C1, C11, C14 y C15.



**Figura 14.** Expresión de las quimeras *in vivo*. Western Blot anti-CAT de las mismas clonas. Las cuatro PRAI quiméricas (C1, C11, C14 y C15) se expresaron fusionadas a CAT.

## 6.2 DISEÑO Y PRUEBA DE OLIGONUCLEÓTIDOS DEGENERADOS PARA LA AMPLIFICACIÓN DE ORTÓLOGOS DISTANTES.

*Obtención de las secuencias y alineamientos múltiples.* Para el diseño de los oligos degenerados se obtuvieron las secuencias de los dominios representantes de CATH v3.0.0 de las 21 superfamilias homólogas con topología barril TIM. A cada uno de los representantes se les realizó un Blastp contra el genoma *E. coli* K12 para elegir el mejor ortólogo. Para seis representantes (de las superfamilias con clave en CATH 3.20.20 terminación 40, 80, 110, 120, 240 y 350 y 3.40.50.280<sup>1</sup>) no se encontraron resultados significativos ( $v.e. \leq 10^{-3}$ ) en *E. coli* K12, por lo cual se seleccionaron otros dominios como representantes de las mismas superfamilias; para tres de estas superfamilias, sin embargo, no se encontraron proteínas significativamente similares en *E. coli* y no se prosiguió trabajando con ellas (3.20.20 terminación 80, 110 y 350). Para tres de los dominios representantes se obtuvieron más de un resultado (pertenecientes a las superfamilias 3.20.20 terminación 10, 100 y 190), por lo cual se eligió al representante en *E. coli* según el mejor valor de expectancia o el mayor porcentaje de identidad en el alineamiento (Tabla 4).

A su vez, para las 18 superfamilias con dominios barril TIM representados en proteínas de *E. coli* se realizaron Blastp contra todos los genomas registrados en NCBI (2005) y bajo el criterio del mejor resultado recíproco (*bidirectional best hit*) se obtuvieron los identificadores de genomas (*gi*) de los resultados significativos y sus secuencias, con las cuales se realizaron alineamientos múltiples para cada grupo de superfamilias (Tabla 4). Bajo el criterio del mejor resultado recíproco, dado un par de genes  $x$  y  $z$  de los genomas  $X$  y  $Z$ , respectivamente, al hacer una búsqueda de ortólogos partiendo de  $x$  en el genoma  $Z$ , en el que  $z$  obtuvo el mejor valor de expectancia y al hacer la búsqueda a partir de  $z$  en el genoma  $X$ ,  $x$  fue el mejor resultado, al ser el mejor resultado recíproco, para  $x$  se elige  $z$ .

En total se obtuvieron 1846 secuencias, con un máximo de 212 secuencias por superfamilia (superfamilia de aldolasas, clave 3.20.20.70) y un mínimo de 36 (superfamilias de alanina racemasas y enolasas, clave 3.20.20 10 y 120 respectivamente). En todos los alineamientos los porcentajes de identidad en las secuencias de aminoácidos son mayores al 30% (Tabla 4). La mayoría de las proteínas obtenidas están anotadas de la misma forma que las proteínas representantes con las cuales se realizó el Blast (*eg.* alanina racemasas), aunque de una a siete

---

<sup>1</sup> En la nueva versión del CATH (v3.1.0) la proteína 1bmt, utilizada como representante para una de las superfamilias en lugar de 1ccw, está clasificada como un dominio tipo Rossmann-fold (arquitectura “emparedeado  $\alpha/\beta/\alpha$  de tres capas”), por lo cual se le presenta con su nueva clave del CATH; la proteína 1req, por otra parte, en la nueva versión está clasificada como la superfamilia 3.20.20.240, a la que 1ccw pertenece, y con la cual se identifica.

proteínas por alineamiento están registradas como proteínas hipotéticas y algunas veces hay proteínas con otra función como indicador, otros barriles TIM y otras veces menos frecuentes, de algún otro dominio.

**Tabla 4.** Superfamilias de barriles TIM, resultados de la búsqueda de ortólogos según el criterio de mejor resultado recíproco (*bidirectional best-hit*) y de alineamientos.

| código CATH | nombre de superfamilia  | PDB representantes | gi de representantes de <i>E. coli</i> | número de secuencias | promedio longitud secuencias | diferencia longitud | % identidad |
|-------------|---|--------------------|--|----------------------|------------------------------|---------------------|-------------|
|             | representantes superfamilias CATH representantes en <i>E.coli</i> |                    |  | 18                   | 304                          | 364                 | 9           |
| 3.20.20.10  | Alanine racemase  | 1bd0A1(s)          | 16129153                               | 36                   | 366                          | 58                  | 39          |
| 3.20.20.20  | Dihydropteroate (DHP) synthetase                                  | 1eyeA0             | 49176323                               | 196                  | 314                          | 626                 | 37          |
| 3.20.20.30  | FMN dependent fluorescent proteins                                | 1lucA0             | 16131052                               | 97                   | 336                          | 297                 | 39          |
| 3.20.20.40  | 7-stranded glycosidases (cellulases)                              | 1ep3A0 (m)         | 49176192                               | 157                  | 349                          | 823                 | 31          |
| 3.20.20.60  | Phosphoenolpyruvate-binding domains                               | 1s2wA0             | 16128316                               | 83                   | 302                          | 187                 | 44          |
| 3.20.20.70  | Aldolase class I  | 1n55A0             | 16131757                               | 212                  | 257                          | 431                 | 41          |
| 3.20.20.80  | Glycosidases  | (s/r)              |  |                      |                              |                     |             |
| 3.20.20.100 | NADP-dependent oxidoreductase                                     | 1us0A0 (s)         | 49176300                               | 108                  | 282                          | 121                 | 44          |
| 3.20.20.105 | tRNA-guanine (tRNA-G) transglycosylase                            | 1r5yA0             | 16128391                               | 209                  | 398                          | 398                 | 44          |
| 3.20.20.110 | Rubisco   | (s/r)              |  |                      |                              |                     |             |
| 3.20.20.120 | Enolase superfamily   | 1r6wA01 (m)        | 16130196                               | 36                   | 337                          | 344                 | 45          |
| 3.20.20.140 | Metal-dependent hydrolases  | 1hzyA0             | 16131257                               | 38                   | 323                          | 536                 | 32          |
| 3.20.20.150 | Divalent-metal-dependent TIM barrel enzymes                       | 1muwA0             | 16131436                               | 52                   | 439                          | 93                  | 54          |
| 3.20.20.190 | Phosphatidylinositol (PI) phosphodiesterase                       | 1o1zA0 (s)         | 16130174                               | 68                   | 351                          | 554                 | 37          |
| 3.20.20.210 | [PDB] Lyase   | 1r3sA0             | 16131827                               | 168                  | 352                          | 408                 | 39          |
| 3.20.20.220 | [PDB] Oxidoreductase  | 1v93A0             | 16131779                               | 102                  | 314                          | 389                 | 40          |
| 3.20.20.240 | [PDB] Isomerase   | 1reqA01            | 16130818                               | 62                   | 672                          | 848                 | 53          |
| 3.40.50.280 | [PDB] Isomerase   | 1bmtA2 (m)         | 16131845                               | 122                  | 1046                         | 1,075               | 40          |
| 3.20.20.300 | [PDB] Hydrolase   | 1ex1A1             | 16130070                               | 59                   | 737                          | 569                 | 34          |
| 3.20.20.330 | [PDB] Transferase   | 1lt8A0             | 16128246                               | 41                   | 364                          | 329                 | 33          |
| 3.20.20.350 | [PDB] Lyase   | (s/r)              |  |                      |                              |                     |             |

(s) familias super-representadas; (m) representante de CATH modificados; (s/r) sin ortólogos en *E. coli* K12; *diferencia longitud*, diferencia entre secuencia más larga y más corta en alineamiento. Este análisis se realizó a partir de CATH v3.0.0.

*Identificación de las zonas de conservación de los alineamientos múltiples.* A partir de los alineamientos múltiples se buscaron regiones de conservación por dos programas distintos: el BMAP (Blocks Multiple Alignment Processor) y MEME (Múltiple *Em* for Motif Elicitation). En el BMAP se realizan bloques –segmentos de una alineación múltiple sin *gaps*–, por identificación de regiones de caracteres no conservados flanqueados por caracteres conservados. BMAP tiene la opción de buscar los oligos degenerados directamente con el programa CODEHOP (Consensus-degenerate hybrid oligonucleotide primers). Los oligos diseñados en CODEHOP contienen todas las posibles secuencias nucleotídicas correspondientes a 3 o 4 aminoácidos conservados en el extremo-3' degenerado y una región constante en el extremo-5' que representa secuencias consenso. En el hipervínculo se permite regular la degeneración máxima de los oligos, la TM teórica –por regular la longitud de los oligos–, la probabilidad mínima permitida para cada nucleótido en un sitio determinado para ser considerado en el algoritmo (precisión) y el código genético a utilizar. El programa MEME identifica motivos –los cuales tampoco pueden incluir *gaps*– a partir del algoritmo de maximización de la expectancia (*EM*, por sus siglas en inglés), por el cual se eligen las ventanas del rango de tamaño predeterminado que posean el mejor resultado en las matrices de probabilidad posición-específicas y los mejores valores de expectancia –es decir, que sean menos probables de ocurrir al azar–.

Para todas las superfamilias se encontraron oligos en CODEHOP, aunque para la mayoría de éstos (13/18) los oligos fueron diseñados a partir de un solo bloque –una región muy cercana dentro de la secuencia–, y por tanto fueron por si solos insuficientes para amplificar un fragmento de gen representativo. No se encontró ninguna relación entre el número de secuencias en el alineamiento, el tamaño de las proteínas en él o su porcentaje de identidad y la producción de oligos por CODEHOP. Por otra parte, para todas las superfamilias se encontraron los motivos de la ocurrencia (de 0 a 1) y del rango de tamaño (de 25 a 50) establecidos en MEME. Todos los motivos presentaron valores de expectancia menores a v.e.  $\leq 10^{-3}$ .

*Diseño de oligonucleótidos degenerados.* Para continuar con el diseño de oligos degenerados se seleccionaron las mismas 4 superfamilias que para las cuales se diseñaron los oligos específicos: las superfamilias 3.20.20.10 (1bd0 como dominio representante), 3.20.20.150 (representante: 1muwA0) 3.20.20.105 (representante: 1r5yA0) y 3.20.20.330 (representante: 1lt8A0). El diseño final de los pares de oligos para estas superfamilias resultó de la combinación de los resultados de CODEHOP y MEME, con algunos cambios posteriores para maximizar los criterios que se enlistan a continuación:

- Tamaño del oligo de preferencia entre 17 a 28 pb;

- %GC en un rango preferencial del 50 al 60%, sin más de 2 bases de guanina o citosina en el extremo-3’;
- TM entre 55 y 80 °C, similar entre ambos oligos;
- Degeneración máxima de 256, de preferencia hacia el extremo-3’ (la degeneración de un sitio se calcula como la suma de las posibles bases que se pueden presentar en él—de 1 a 4-, mientras que la degeneración del oligo es el producto de la degeneración de todos sus sitios);
- $\Delta G$  de homo o heterodimerización  $\geq -3$  kcal/mol o menos de 3 *pb* complementarias;
- El producto amplificado por PCR tamaño del amplificado debe ser de preferencia entre 350 y 600 *pb*.
- Los oligos deben de amplificar la mayor parte de la secuencia de cada proteína, es decir, estar en regiones del alineamiento múltiple lo suficientemente separadas, maximizando la representatividad de la secuencia amplificada.
- En caso de tener varias opciones, escoger aquellos pares de oligos con los cuales se tengan mejores resultados en un Blastn contra todos los genomas bacterianos de NCBI (2005) (búsqueda por 5’-secuencia\_oligo\_1-3’—NN—5’-secuencia\_oligo\_2-3’).

Los tres oligos detectados en CODEHOP para la superfamilia del dominio 1bd0 fueron complemento del primer bloque en el alineamiento y con la máxima degeneración permitida (256), por lo cual se diseñaron los oligos a partir de los motivos del inicio y final de las secuencia generados en MEME. En el oligo *forward*, se agregaron dos bases en el extremo 5’ para aumentar el %GC y el tamaño del mismo y se eliminaron dos degeneraciones (S→G y R→A); en el oligo reverso se modificaron seis degeneraciones (V→S, Y→C, R→G, V→R, M→C y R→G). En el caso de la superfamilia del dominio 1muwA, los dos oligos generados por CODEHOP resultaron útiles, por lo cual se sintetizaron tal con 4 modificaciones en el *forward* y 5 en el reverso. El oligo *forward* de la superfamilia del dominio 1lt80 se extrajo a partir del motivo generado por MEME tras dos eliminaciones de degeneración (S→G y S→C), mientras que el oligo reverso se extrajo del resultado de CODEHOP eliminando la degeneración. Por último, el oligo *forward* de la superfamilia con representante 1r5y se extrajo se extrajo del motivo localizado al final del alineamiento generado por MEME con tres modificaciones y el reverso se obtuvo de CODEHOP directamente, aunque se redujeron dos degeneraciones. En el Apéndice 1 se muestran los oligos finales, remarcando los sitios en los cuales se hicieron las modificaciones y en la Tabla 5 se presentan las características de las condiciones que se buscaron maximizar.

**Tabla 5.** Características de los oligos degenerados diseñados para amplificar ortólogos lejanos de la PRAI<sub>Ec</sub> a partir de una muestra ambiental.

| Oligo      | Origen | Longitud (pb) | %GC  | <i>T<sub>m</sub></i> | deg | ΔG homodímero/heterodímero | Pb amp/% amp |
|------------|--------|---------------|------|----------------------|-----|----------------------------|--------------|
| 1bd0 5'-3' | MEME   | 17            | 58.9 | 60.8                 | 32  | 3.61                       |              |
| 1bd0 3'-5' | MEME   | 19            | 49.1 | 52                   | 48  | 7.26/8.75                  | 837/77.5     |
| 1muw 5'-3' | BMAP   | 29            | 56.9 | 64.1                 | 96  | 12.46                      |              |
| 1muw 3'-5' | BMAP   | 31            | 45.2 | 62                   | 128 | 10.46/7.87                 | 622/46.97    |
| 1r5y 5'-3' | MEME   | 32            | 42.2 | 61.4                 | 16  | 12.9                       |              |
| 1r5y 3'-5' | BMAP   | 17            | 44.1 | 47.2                 | 16  | 5.38/5.38                  | 600/54.49    |
| 1lt8 5'-3' | MEME   | 17            | 58.8 | 56.1                 | 64  | 5.37                       |              |
| 1lt8 3'-5' | BMAP   | 22            | 40.9 | 50                   | 0   | 3.91/8.26                  | 604/64.66    |

*pb*, pares de bases; %GC, porcentaje de guaninas y citocinas; *T<sub>m</sub>*, temperatura de hibridación; ΔG homodimerización/ΔG heterodimerización; *pb amp*/*% amp*, *pb* esperadas en PCR de *E. coli*/ porcentaje del gen blanco parálogo de *E. coli* amplificado.

*Resultados de amplificación con los oligonucleótidos degenerados en E. coli K12.* Los cuatro pares de oligos se probaron en *E. coli* K12 por PCR con gradientes de temperatura de hibridación. En todas las reacciones se observaron bandas del tamaño esperado del producto blanco y al menos cuatro productos inespecíficos y dimerización de oligos. No se encontró para ningún par de oligos una *T<sub>m</sub>* en la cual se amplificara únicamente el producto del tamaño esperado, aunque de ésta si depende que se amplifique del todo o no. Se clonaron y secuenciaron tres de los productos de los oligos 1muw y uno de los oligos 1bd0. Uno de los productos se identificó como barril TIM: un fragmento (de ~200 *pb*) de un gen parálogo a la xilosa isomerasa a la cual se le aplicó el criterio del mejor resultado recíproco con el resto de los genomas; la mayoría de los demás dominios identificados corresponden a arquitecturas del tipo de α/β/α sándwich (Tabla 6).

**6.2.1 Conclusión y perspectivas.** Se logró amplificar en *E. coli* un barril TIM a partir de un par de oligos degenerados diseñado por el reconocimiento de áreas de conservación hacia los extremos de las secuencias de una superfamilia de barril TIM. Los oligos no fueron diseñados para *E. coli* particularmente, por lo que estos resultados parecen apoyar la metodología de su diseño y el sistema planteado en este trabajo.

**Tabla 6.** Identificación de algunos fragmentos amplificados por los oligos degenerados.

| oligo | Pb amp | v.e.   | Amp/<br>gen | nombre proteína   | gi       | longitud gen<br>(pb) | arquitectura                       | topología                               |
|-------|--------|--------|-------------|---|----------|----------------------|------------------------------------|---|
| 1muw  | 170    | 3e-53  | 170         | Xilosa isomerasa  | 1789988  | 3075<br>(2895-3065)  | Barril $\alpha/\beta$              | Barril Tim                              |
| 1muw  | 527    | 0      | 313         | lipoproteína<br>hipotética<br>precursor yfhM                          | 1788868  | 4959<br>(4646-4959)  | -                                  | -                                       |
|       |        |        | 214         | 3-mercapto-<br>piruvato sulfo-<br>transferasa                         | 87082121 | 846 (1-214)          | $\alpha\beta$ Sándwich             | rodonasa<br>oxidada                     |
| 1muw  | 295    | 1e-141 | 295         | Guanosin-3',5'-<br>bis difosfato 3'-<br>pirofosfo-<br>hidrolasa*      | 1790082  | 2106<br>(1431-1726)  | Proteínas de $\alpha$ -<br>hélices | Dominio- HD                             |
|       |        |        |             |   |          |                      | Proteínas ( $\alpha+\beta$ )       | Nucleotidil<br>transferasa              |
| 1bd0  | 610    | 1e-166 | 41          | Precursor<br>periplásmico de<br>unión a histidina                     | 1788648  | 783 (742-783)        | $\alpha\beta$ Sándwich             | Proteína de<br>unión D-<br>Maltodextrin |
|       |        |        | 569         | Precursor<br>periplásmico de<br>unión a Lisina-<br>arginina-ornitina- | 1788649  | 780 (1-569)          | $\alpha\beta$ Sándwich             | Proteína de<br>unión D-<br>Maltodextrin |

*pb*, pares de bases; *Pb amp*, pares de bases amplificadas; *v.e.* valor de expectancia de un *Blastn* vs. *E. coli* con el fragmento secuenciado. *Amp/gen*, pares de bases amplificadas por gen; longitud gen, pares de bases del gen original -en paréntesis, localización del fragmento amplificado-. \*Clasificación según SCOP.

La obtención de productos de barril TIM o específicos no es trivial. Los oligos degenerados amplificaron de 4 a 8 productos inespecíficos, mientras que con los oligos específicos se obtuvieron, por par, alrededor de 3 productos inespecíficos. Disminuir la degeneración de los oligos, por lo tanto, no necesariamente disminuirá o eliminará su inespecificidad; sin embargo, se puede buscar maximizar la amplificación del producto esperado por PCRs con gradiente de temperatura de hibridización y recolección por purificación por geles de agarosa. Los productos inespecíficos, por otra parte, pueden ser fragmentos de los genes blanco, homólogos de los mismos, secuencias con cierta similitud u otro tipo de fragmentos de DNA, que, en caso de que sean éstos los que recombinan, no estropean el experimento. Los resultados de esta sección, por otro lado, son únicamente la prueba de los oligos en un solo genoma; no se descarta que en otros genomas haya más o menos especificidad con el mismo par de oligos.

La depuración de las secuencias que se incluyen en la alineación múltiple puede ser fundamental en la optimización del diseño de los oligos. El criterio del mejor resultado recíproco

asegura la inclusión de secuencias con el mejor valor de expectancia, sin embargo no implica que éstas sean ortólogas. Aunque una rápida comparación de los identificadores de las enzimas obtenidas pueda ser informativa de su homología, casos como el presentado entre las superfamilias con 1bmt y 1lt8 como representantes –metil transferasas mayoritariamente pero de dominio Rossmann fold y barril TIM, respectivamente- ejemplifican que un análisis más profundo de las secuencias incluidas en el análisis puede mejorar la calidad de la exploración desde el inicio. Por otro lado, tanto los bloques como los motivos tienen la ventaja de eliminar en el análisis los distintos dominios que pueden estar río abajo o río arriba del dominio de interés; sin embargo, también tienden a homogenizar los pesos de las secuencias en el análisis –los bloques por normalizar el peso de cada secuencia y los motivos por igualarlos-. Por lo anterior, se recomienda regular manualmente los pesos según la naturaleza de la muestra ambiental que se desee utilizar y eliminar secuencias muy lejanas, de tamaños desmesurados o redundantes.

## 7. CONCLUSIÓN

Los resultados de trabajos anteriores de complementación de proteínas comentados en los antecedentes –como la complementación por fragmentos de DNA genómico de la CspA (Riechmann 2000), de la lisozima de huevo de gallina (Riechmann 2006) y de la PRAI<sub>Ec</sub> (Sánchez 2006)- ejemplifican la plasticidad de los plegamientos concedida por el barajeo de módulos. Ante tal plasticidad, uno podría pensar que cualquier segmento de una secuencia de cierto genoma tiene potencial para producir dominios plegados al complementarse con otros segmentos del mismo genoma. Racionalizando lo anterior surge la siguiente pregunta: ¿qué tipo de secuencias se deberían de utilizar para optimizar la complementación?

En contraposición a una librería generada a partir de fragmentos de DNA genómico, la gama de secuencias proveniente de fragmentos de genes amplificados por PCR es muy limitada, por lo cual se podría predecir una drástica reducción en las probabilidades de complementación. Sin embargo, mientras más pares de oligonucleótidos y más genomas se incluyan en trabajos de ésta índole, menor reducción de la gama de secuencias iniciales. Por otra parte, la utilización de homólogos lejanos de varias pozas génicas, al haber tenido una historia en común y ser genes codificantes, pueden restringir la inclusión de mutaciones deletéreas, aumentando así la probabilidad de éxito de las complementaciones y fomentando una extensa exploración del área de secuencia.

En este trabajo no se cuantificó ni comparó el éxito de las recombinaciones no homólogas a partir de secuencias al azar y homólogos distantes. Sin embargo, al haber logrado la generación de cuatro proteínas quiméricas plegadas a partir de cuatro pares de oligonucleótidos, un solo genoma y un banco de tamaño muy reducido, se puede intuir que el método aquí planteado sí puede aumentar las probabilidades del éxito de las recombinaciones no homólogas. Por otro lado, el método es asequible, optimizable y competente para conseguir proteínas quiméricas. Al haber utilizado un tipo de barril TIM, el método planteado se puede aplicar en proteínas con el mismo dominio o con otro distinto, considerando que entre más diversos éstos sean y más modulares, mejores resultados se pueden obtener.

## 8. APÉNDICES

### Apéndice 1: Lista de oligonucleótidos utilizados a lo largo del trabajo.

| oligo   | Identificador | clave     | TM (°C) | SECUENCIA  |
|---|---------------|-----------|---------|--|
| Oligos específicos para parálogos de PRAI <sub>Ec</sub>                               |               |           |         |  |
| alr 5'-3'   | 1bd_ecoli_for | 2006-832  | 62.56   | AAT GGT TGC GGT GGT GAA                          |
| alr 3'-5'   | 1bd_ecoli_rev | 2006-833  | 64      | TCG TCA TTT CAG CGA TAC GTT C                    |
| xyl 5'-3'   | 1mu_ecoli_for | 2006-834  | 66.55   | CCG TTA GCA TTC CGT CAC TAC A                    |
| xyl 3'-5'   | 1mu_ecoli_rev | 2006-835  | 76      | TTG CTC CTG GCG ACC ACT CTG ATG                  |
| tRg 5'-3'   | 1ry_ecoli_for | 2006-836  | 64.61   | TTG AAC TGG ACA CCA CCG A                        |
| tRg 3'-5'   | 1ry_ecoli_rev | 2006-837  | 69.46   | CTT GCG TAA ACC CGC CAT CAA ACG                  |
| hmt 5'-3'   | 1lt_ecoli_for | 2006-838  | 66.45   | GAT ATC CTG CTG CTG GAT GG                       |
| hmt 3'-5'   | 1lt_ecoli_rev | 2006-839  | 70      | CGA TAT CCG CAG GCG TGG TGC                      |
| Oligos degenerados  |               |           |         |  |
| 1bd0 5'-3'  | 1bd_5'_3'     | 2005-4174 | 64.92   | <b>GKG KCG GTD GTR AAR GC</b>                    |
| 1bd0 3'-5'  | 1bd_3'_5'     | 2005-4175 | 56      | <b>RCR CAR AGH AYC TCA TAG S</b>                 |
| 1muw 5'-3'  | 1muw_5_3      | 2005-2865 | 74      | GGA GAG AAC TAC GTT TTC TGG GGN GGN <b>MG</b>    |
| 1muw 3'-5'  | 1muw_3_5      | 2005-2866 | 70.84   | CGT ATT GAT GTT TCG TCG GTT <b>CNW KNG</b> GYT T |
| 1r5y 5'-3'  | 1r5y_5_3_2    | 2006-57   | 59.92   | GGA GGT <b>TYT CAR GTN</b> TT                    |
| 1r5y 3'-5'  | 1r5y_3_5      | 2006      | 60      | CGG GTA GGC ATT ACG CAG <b>TCR AAC ATR</b> TC    |
| 1lt8 5'-3'  | 1lt85_3_      | 2006-94   | 54      | <b>MTG GGC ACC GAG TTK</b> GA                    |
| 1lt8 3'-5'  | 1lt83_5_      | 2006-95   | 63.45   | CCG AAC ATA TAC AGG ATA TAG C                    |
| Oligos para amplificar plásmido PT4-Prai-CAT (Escobar 2001; Farías 2005; Morán 2003)  |               |           |         |  |
|   | pT4           | 2004-237  | 63.26   | GAC-ATA-TAA-ACG-GTT-CTG-GCA                      |
|   | NdePRAI       | 2003-3034 | 64.53   | ATA-TAT-ACA-TAT-GCT-GGA-GAA-TAA-GGT-ATG          |
|   | BamPRAIx      | 2004-270  | 75.12   | GCG GAT CCG TAT GCG CGC AGC GTC                  |
|   | CAT100        | 2002-2880 | 64.75   | CTG-AAC-GGT-CTG-GTT-ATA-GG                       |
|   | pT4-2         | 2004-238  | 63.26   | GCC-AGT-GAA-TTG-TAA-TAC-GAC                      |
| Oligos universales para amplificar genes Ribosomales 16s (Baker <i>et al.</i> , 2003) |               |           |         |  |
|   | 27revisado    | 2005-3758 | 63.9    | AGA GTT TGA TCM TGG CTC AG                       |
|   | 1492revis     | 2005-3759 | 64.23   | TAC GGY TAC CTT GTT ACG ACT T                    |

### Apéndice 2: Reparación

- Reparación por T4-DNA polimerasa (T4 pol) y Klenow.

|                              |                  |
|------------------------------|------------------|
| Inserto                      | de 800 ng a 1 µg |
| Buffer T4 pol (Biolabs, 10x) | 3 µl             |
| Klenow (Biolabs, 5 u/µl)     | 4 µl             |
| T4 pol (Biolabs, 3 u/µl)     | 4 µl             |
| dNTPs (2.5 mM)               | 5 µl             |
| H <sub>2</sub> O             | Vf = 30 µl       |

Programa en el termociclador: 12 °C 30 min favoreciendo la actividad de T4 DNA polimerasa (T4 pol), 25 °C 30 min, temperatura óptima de Klenow y 70 °C 10 min, para desactivar Klenow. Vf, volumen final de la reacción.

- Reparación por exonucleasa Mung-Bean (probado solo en el vector).

|  |                  |
|--|------------------|
| Vector digerido                          | de 800 ng a 1 µg |
| Buffer 2 (Biolabs, 10x)                  | 10 µl            |
| exonucleasa Mung-Bean (Biolabs, 10 u/ml) | 0.5 µl           |
| H <sub>2</sub> O                         | Vf = 100 µl      |

1. Incubar 60 min a 25 °C y parar la reacción con 12 µl de EDTA 100 mM.
2. Llevar volumen final a 110 µl con H<sub>2</sub>O..

De cualquier manera se puede solamente concentrar o directamente seguir con el procedimiento si se utiliza en la desfosfatación el mismo Buffer de reacción. Si se desea proceder a la ligación de inmediato para utilizar la muestra como control, se deberá limpiar con Roche PCR purification Kit, fenol-cloroformo o butanol-etanol.

**Apéndice 3:** Digestión doble de PT4-PRAI-CAT con BanI y MslII

|                         |            |
|-------------------------|------------|
| Vector                  | 1 µg       |
| Buffer 2 (Biolabs, 10x) | 1.5 µl     |
| BanI (Biolabs, 20 u/ml) | 0.5 µl     |
| MslII (Biolabs, 5 u/ml) | 1 µl       |
| H <sub>2</sub> O        | Vf = 15 µl |

Incubar *ON* a 37 °C, sabiendo que las enzimas no tienen alta actividad estrella. Purificar por banda para bajar la probabilidad de trabajar con plásmido superenrollado.

**Apéndice 4:** Desfosfatación de los extremos libres de PT4-PRAI-CAT.

|                            |                  |
|----------------------------|------------------|
| Plásmido reparado y limpio | de 800 ng a 1 µg |
| Buffer 3                   | 5.0 µl           |
| CIP (Biolabs, 10 u/µl)     | 0.5 µl           |
| H <sub>2</sub> O           | Vf 50 µl         |

1. Incubar 30 min a 37 °C
2. 30 min a 67 °C
3. Agregar 6.5 µl EGTA (250 mM) y llevar al Vf = 80 (*ie.* 23.5 µl H<sub>2</sub>O 4d)
4. 30 min a 67 °C (de nuevo)
5. Limpiar con Roche PCR purification Kit, fenol-cloroformo (aunque residuos de estas sustancias pueden inhibir la ligación) o butanol-etanol.
6. Resuspender en 30 µl con agua tetradestilada.

**Apéndice 5:** Ligación para la generación de los bancos.

|                                     |               |
|-------------------------------------|---------------|
| Inserto:Vector::10:1                | de 250-500 ng |
| Buffer T4-DNA ligasa (Biolabs, 10x) | 2 µl          |
| T4-DNA ligasa (Biolabs, 10 u/ml)    | 1 µl          |
| BSA 10x                             | 2 µl          |
| H <sub>2</sub> O                    | Vf = 20 µl    |

Incubar *ON* a 16 °C, de preferencia en tubos de 500 µl, conocidos por su poca absorción de DNA. Limpiar la muestra con butanol y precipitar con etanol al 70%.

## 9. REFERENCIAS

- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W y Lipman DJ. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25.
- Akanuma S y Yamagishi A. (2005). Identification and characterization of key substructures involved in the early folding events of a (beta/alpha)<sub>8</sub>-barrel protein as studied by experimental and computational methods. *J Mol Biol.* 353(5).
- Apic G, Gough J y Teichmann SA. (2001). Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol.* 310(2).
- Arnold H. (2006). Fancy footwork in the sequence space shuffle. *Nature Biotechnology* 24.
- Bacher J, Reiss B y Ellington A. (2002). Anticipatory evolution and DNA shuffling. *Genome Biology* 3.
- Bono S, Riechmann L, Girard E, Williams RL, y Winter G. (2005). A segment of cold shock protein directs the folding of a combinatorial protein. *PNAS* 102(5).
- Bogarad L y Deem M. (1999). A hierarchical approach to protein molecular evolution. *Proc. Natl. Acad. Sci. USA* 96.
- Bornberg-Bauer E, Beaussart F, Kummerfeld SK, Teichmann SA y Weiner J. (2005). The evolution of domain arrangements in proteins and interaction networks. *Cell Mol Life Sci.* 62(4).
- Bornberg-Bauer E y Sun Chan A. (1999). Modeling evolutionary landscapes: Mutational stability, topology, and superfunnels in sequence space. *Proc. Natl. Acad. Sci. USA* 96.
- Brannigan JA y Wilkinson AJ. (2002). Protein engineering 20 years on. *Nat Rev Mol Cell Biol.* 3(12).
- Bryson K, McGuffin LJ, Marsden RL, Ward JJ, Sodhi JS y Jones DT. (2005). Protein structure prediction servers at University College London. *Nucl. Acids Res.* 33.
- Boubakri H, Beuf M, Simonet P y Vogel TM. (2006). Development of metagenomic DNA shuffling for the construction of a xenobiotic gene. *Gene* 375.
- Chung SY y Subbiah S. (1996). A structural explanation for the twilight zone of protein sequence homology. *Structure* 4(10).
- Copley RR y Bork P. (2000). Homology among (alpha/beta)<sub>8</sub> barrels: implications for the evolution of metabolic pathways. *J Mol Biol.* 303(4).
- Creighton TE. (ed; 1992). Protein folding. W.H. Freeman, New York.
- Cui y Wong WH, Bornberg-Bauer W y Chan HS. (2002). Recombinatoric exploration of novel folded structures: A heteropolymer-based model of protein evolutionary landscapes. *PNAS* 99(2).
- Doi N y Yanagawa H. (1999). Design of generic biosensors based on green fluorescent proteins with allosteric sites by directed evolution. *FEBS Letters* 453(3).
- Drummond DA, Silberg JJ, Meyer MM, Wilke CO y Arnold FH. (2005). On the conservative nature of intragenic recombination. *PNAS* 102(15).
- Eddy SR. (1998). Profile hidden Markov models. *Bioinformatics* 14.
- Finke JM y Onuchic JN. (2005). Equilibrium and kinetic folding pathways of a TIM barrel with a funneled energy landscape. *Biophys J.* 89(1).
- Frankel E. (1980). DNA el proceso de la vida. Ed. Siglo Veintiuno Editores, SA. México, D.F. 10º Edición. Capítulo 3. La química de la vida, pag. 29-41.
- Futuyma D. (1998). Evolutionary biology. Third edition. Sinauer Associates, Inc., Sunderland, Massachusetts. 763 pp. Gerlt JA y Raushel FM. (2003). Evolution of function in (alpha/beta)<sub>8</sub>-barrel enzymes. *Curr Opin Chem Biol.* 7(2).
- Goraj K, Renard A y Martial JA. (1990). Synthesis, purification and initial structural characterization of octarellin, a de novo polypeptide modeled on the alpha/beta-barrel proteins. *Protein Eng.* 3(4).
- Grishin NV. (2001). Fold change in evolution of protein structures. *J Struct Biol.* 134(2-3).
- Gromiha MM, Pujadas G, Magyar C, Selvaraj S y Simon I. (2004). Locating the stabilizing residues in (alpha/beta)<sub>8</sub> barrel proteins based on hydrophobicity, long-range interactions, and sequence conservation. *Proteins* 55(2).
- Höcker B, Claren J y Sterner R. (2004). Mimicking enzyme evolution by generating new (beta/alpha)<sub>8</sub>-barrels from ((alpha/beta)<sub>4</sub>-half-barrels. *Proc Natl Acad Sci.* 101(47).
- Houbrechts A, Moreau B, Abagyan R, Mainfroid V, Preaux G, Lamproye A, Poncin A, Goormaghtigh E, Ruyschaert JM, Martial JA. (1995). Second-generation octarellins: two new de novo (beta/alpha)<sub>8</sub> polypeptides designed for investigating the influence of beta-residue packing on the alpha/beta-barrel structure stability. *Protein Eng.* 8(3).
- Janin J y Wodak SJ. (1983). Structural domains in proteins and their role in the dynamics of protein function. *Prog Biophys Mol Biol.* 42(1).
- Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M y Hirakawa M. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* 34.

- Kinch L y Grishin N. (2002). Evolution of protein structures and functions. *Current Opinion in Structural Biology* 12.
- Koonin EV, Wolf YI y Karev GP. (2002). The structure of the protein universe and genome evolution. *Nature* 420(6912).
- Lehninger AL. (1982). Principios de Bioquímica. Omega. Barcelona, España
- Livesay DR y La D. (2005). The evolutionary origins and catalytic importance of conserved electrostatic networks within TIM-barrel proteins. *Protein Sci.* 14(5).
- Loughlin T, Patrick WM y Matsumura I. (2006). Natural history as a predictor of protein evolvability. *Protein Engineering, Design & Selection* 19(10).
- Luger K, Szadkowski H y Kirschner K. (1990). An 8-fold beta alpha barrel protein with redundant folding possibilities. *Protein Eng.* 3(4).
- Luger K, Hommel U, Herold M, Hofsteenge J y Kirschner K. (1989). Correct folding of circularly permuted variants of a beta alpha barrel enzyme *in vivo*. *Science* 243 (4888).
- Lupas A, Ponting C y Russell R. (2001). On the Evolution of Protein Folds: Are Similar Motifs in Different Protein Folds the Result of Convergence, Insertion, or Relics of an Ancient Peptide World? *Journal of Structural Biology* 134.
- Marsden RL, Ranea JA, Sillero A, Redfern O, Yeats C, Maibaum M, Lee D, Addou S, Reeves GA, Dallman TJ y Orengo CA. (2006). Exploiting protein structure data to explore the evolution of protein function and biological complexity. *Philos Trans R Soc Lond B Biol Sci.* 361(1467).
- Maxwell KL, Mittermaier AK, Forman-Kay JD y Davidson AR. (1999). A simple *in vivo* assay for increased protein solubility. *Protein Sci.* 8.
- Moore G, Maranas C, Lutz S y Benkovic S. (2001). Predicting crossover generation in DNA shuffling. *PNAS* 98(6).
- Murzin A, Brenner S, Hubbard T y Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247.
- Moore GL, Maranas CD, Lutz S y Benkovic SJ. (2001). Predicting crossover generation in DNA shuffling. *Proc Natl Acad Sci U S A.* 98(6).
- Nagano N, Orengo CA y Thornton JM. (2002). One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J Mol Biol.* 321(5).
- Orengo C, Michie A, Jones S, Jones D, Swindells M y Thornton J. (1997). CATH- A Hierarchic Classification of Protein Domain Structures. *Structure* 5(8).
- Ostermeier M, Shim JH y Benkovic SJ. (1999). A combinatorial approach to hybrid enzymes independent of DNA homology. *Nat Biotechnol.* 17(12).
- Pal C, Papp B y Lercher MJ. (2006). An integrated view of protein evolution. *Nat Rev Genet* 7(5).
- Patthy L. (2003). Modular assembly of genes and the evolution of new functions. *Genetica* 118(2-3).
- Pirun M, Babnigg G y Stevens FJ.(2005). Template-based recognition of protein fold within the midnight and twilight zones of protein sequence similarity. *J Mol Recognit.* 18(3).
- Peimbert M. (2003). Migración catalítica en la superfamilia de las b-lactamasas: obtención de la actividad  $\beta$ -lactamasa en una DD-peptidasa. Tesis doctoral, Instituto de Biotecnología, UNAM, Cuernavaca, Morelos.
- Ponting CP y Russell RR.(2002). The natural history of protein domains. *Annu Rev Biophys Biomol Struct* 31.
- Priestle JP, Grutter MG, White JL, Vincent MG, Kania M, Wilson E, Jardetzky TS, Kirschner K y Jansonius JN. (1987). Three-dimensional structure of the bifunctional enzyme N-(5'-phosphoribosyl)anthranilate isomerase-indole-3-glycerol-phosphate synthase from *Escherichia coli*. *Proc Natl Acad Sci* 84(16).
- Reardon D y Farber GK. (1995). The structure and evolution of alpha/beta barrel proteins. *FASEB J.* 10(1).
- Riechmann L y Winter G. (2006). Early protein evolution: building domains from ligand-binding polypeptide segments. *J Mol Biol.* 363(2).
- Riechmann L y Winter G. (2000). Novel folded protein domains generated by combinatorial shuffling of polypeptide segments. *PNAS* 97(18).
- Rose TM, Schultz ER, Henikoff JG, Pietrokovski S, McCallum CM y Henikoff S. (2002). Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly-related sequences. *Nucleic Acids Research*, 26(7).
- Salem GM, Hutchinson EG, Orengo CA y Thornton JM. (1999). Correlation of Observed Fold Frequency with the occurrence of Local Structural Motifs. *J. Mol. Biol.* 287.
- Sambrook J, Fritsch EF y Maniatis T. (1989). Molecular Cloning. A laboratory manual. 2<sup>a</sup> ed. Cold Spring Harbor Laboratory Press (3 volúmenes).

- Sánchez L. (2006). Obtención de proteínas plegadas por complementación con secuencias genómicas. Tesis profesional. Facultad de Ciencias, UNAM, México D.F.
- Shukla A y Guptasarma P. (2004). Folding of beta/alpha-unit scrambled forms of *S. cerevisiae* triosephosphate isomerase: Evidence for autonomy of substructure formation and plasticity of hydrophobic and hydrogen bonding interactions in core of ( $\beta\alpha$ )<sub>8</sub>-barrel. *Proteins* 55(3).
- Söding J y Lupas A. (2003). More than the sum of their parts: on the evolution of proteins from peptides. *BioEssays* 25.
- Soberón X, Fuentes-Gallego P y Saab-Rincón G. (2004). In vivo fragment complementation of a (beta/alpha)<sub>8</sub> barrel protein: generation of variability by recombination. *FEBS Lett.* 560.
- Souza V, Castillo A, Rocha M, Sandner L, Silva C y Eguiarte L. (2001). Ecología evolutiva de *Escherichia coli*. *Interciencia* [online] 26(10).
- Timothy L y Michael G. (1998). Methods and statistics for combining motif match scores. *Journal of Computational Biology*, 5.
- Tobin MB, Gustafsson C, Huisman GW. (2000). Directed evolution: the 'rational' basis for 'irrational' design. *Curr Opin Struct Biol.* 10(4).
- Tsuji T, Onimaru M y Yanagawa H. (2006). Towards the creation of novel proteins by block shuffling. *Comb Chem High Throughput Screen* 9(4).
- Vogel C y Morea V. (2006). Duplication, divergence and formation of novel protein topologies. *Bioessays* 28(10).
- Voigt CA, Martinez C, Wang ZG, Mayo SL y Arnold FH. (2002). Protein building blocks preserved by recombination. *Nature Structural Biology* 9.
- Yuan L, Kurek I, English J y Keenan R. (2005). Laboratory-directed protein evolution. *Microbiol Mol Biol Rev.* 69(3).
- Walden H, Taylor GL, Lorentzen E, Pohl E, Lilie H, Schramm A, Knura T, Stubbe K, Tjaden B y Hensel R. (2004). Structure and function of a regulated archaeal triosephosphate isomerase adapted to high temperature. *J Mol Biol.* 342(3).
- Watters AL y Baker D. (2004). Searching for folded proteins in vitro and in silico. *Eur J Biochem.* 271(9).
- Wierenga RK. (2001). The TIM-barrel fold: a versatile framework for efficient enzymes. *EBS Lett.* 492(3).
- Wise EL y Rayment I. (2004). Understanding the importance of protein structure to nature's routes for divergent evolution in TIM barrel enzymes. *Acc Chem Res.* 37(3).