00322

UNIVESIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

143

"ANÁLISIS DE LAS SECUENCIAS DE PROTEÍNAS QUE ADOPTAN EL PLEGAMIENTO $(\beta/\alpha)_8$ BARREL"

T E S I S

QUE PARA OBTENER EL TÍTULO DE :

B I Ó L O G A

P R E S E N T A :

LETICIA ORTEGA RUBIO

DIRECTOR DE TESIS: LORENZO SEGOVIA FORCELLA

2003

A

PAGINACIÓN DISCONTINUA

DRA. MARÍA DE LOURDES ESTEVA PERALTA
Jefa de la División de Estudios Profesionales de la
Facultad de Ciencias
Presente

Comunicamos a usted que hemos revisado el trabajo escrito:

Análisis de las secuencias de proteínas que adoptan el plegamiento (alfa/beta)8 barrel.

realizado por Leticia O rtega Rubio

con número de cuenta 9417164-6 , quién cubrió los créditos de la carrera de Biología

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis
Propietario Dr. Lorenzo Segovia Forcella

Propietario Dr. Ernesto Pérez Rueda

Propietario Biol. Luis José Delaye Arredondo

Suplente Dr. Víctor Manuel Valdés López

Suplente Dr. Arturo Carlos II Becerra Bracho

FACULTAD DE CIENCIAS
UNIDAD DE ENSEÑANZA DE BIOLOGIA

Consejo Departamental de Biología

M. en C. Juan Manuel Rodríguez Chávez

B

AGRADECIMIENTOS

La primer persona a la cual quiero agradecer es a mi director de tesis, el Dr. Lorenzo Segovia, por haberme dado todas las facilidades para hacer la tesis, por haber sido tan accesible y obviamente por haberme adentrado en el mundo de la bioinformática.

A Javi, por haberme enseñado tantas cosas, por jalarme las orejas cuando no hacía bien las cosas, por ayudarme siempre que se lo pedía y por ser mi amigo, tu ayuda me es invaluable, gracias.

Quiero también dar las gracias a la familia Huerta Hernández, por haberme ayudado siempre en todo.

A todos mis amigos, Noemí, Irma, Rocío, Armando, Arturo, sin su apoyo, moral y muchas veces económico, nunca hubiera podido terminar. Perdón si omito a alguien, pero saben que de todas maneras los quiero.

Por supuesto, a mi madre, por darme la vida y por ser tan linda conmigo siempre.

C

RESUMEN

Hasta 1990 se pensaba que enzimas estructuralmente relacionadas, catalizaban reacciones químicas idénticas, quizá con sustratos diferentes, es decir, debía existir un plegamiento diferente para cada tipo de reacción química, sin embargo, ese año, se vio que las enzimas mandelato racemasa y la muconato lactonizante adoptan el mismo plegamiento, el de barril TIM y catalizan reacciones diferentes, lo cual atrajo inmediatamente la atención hacia ese plegamiento. A partir de esa fecha se han registrado más enzimas que adoptan ese plegamiento y llevan a cabo reacciones catalíticas muy variadas. A la fecha hay 26 súper familias de barril TIM y llevan a cabo cinco de las seis clases primarias de actividades enzimáticas definidas por la Comisión Enzimática. La pregunta que surge, es si todas las enzimas que adoptan el plegamiento barril TIM son producto de evolución divergente. Para responder a esta pregunta, en este trabajo se analizaron secuencias de las 26 súper familias de barril TIM, tomadas de la base de datos SCOP, para realizar una búsqueda en la base de datos No Redundante del NCBI, empleando el programa PSI-Blast. Se encontraron homologías entre 25 de las 26 súper familias.

ÍNDICE

E

1. Introducción

Si asumimos que el tamaño promedio de una proteína es de 300 aminoácidos, puede haber 20^{300} diferentes secuencias. Obviamente, solo una minúscula parte del espacio potencial de secuencia está poblado por proteínas reales. Por ejemplo, asumiendo que haya 30 millones de especies en la Tierra y que el genoma de cada especie consista de aproximadamente 5 000 genes (un número intermedio entre procariontes y eucariontes) pueden haber 5×10^{10} secuencias de proteínas. Aunque esta cantidad es muy pequeña comparada con el vasto espacio de secuencia, resulta de varios órdenes de magnitud más grande que aquéllas contenidas en las bases de datos hoy en día. Sin embargo, a pesar de estas estimaciones, se ha encontrado que 28% de las entradas en la base de datos no redundante tienen al menos 25% de identidad con una de las entradas de la base de datos estructural[1], lo cual significa que, al menos, una cuarta parte de las secuencias de proteínas conocidas pertenecen a familias para las cuales hay una estructura. Hasta abril 1991[1] se encontraban 83 plegamientos diferentes y para la gran mayoría de ellos, también se encontraban semejanzas en el nivel de secuencia, funcionales y en las orientaciones de las hojas beta y en las hélices alfa, que apoyan sólidamente la idea de que provienen de un ancestro común. Actualmente, existen alrededor de 927 plegamientos distintos[2], resulta claro, a partir de la existencia e identificación de proteínas y genes homólogos, que la población del universo proteico no está distribuida al azar. Sin embargo, para extraer información a partir de esta distribución, se necesita explorar exhaustivamente el universo proteico, lo cual es posible dentro del contexto de una taxonomía jerárquica de las proteínas.

La función biológica de una proteína depende de la forma que adopte en el espacio, es decir de su estructura tridimensional y de algunos aminoácidos clave. Una proteína sólo es activa cuando adopta una estructura determinada. Hasta el momento, se han depositado 19, 379 estructuras en el Banco de Datos de Proteínas (PDB) lo que hace evidente la necesidad de clasificarlas de acuerdo con un criterio homogéneo, pero sobre todo filogenético. A mediados de la década de 1990, Murzin[3] y sus colegas, desarrollaron una taxonomía elaborada y coherente. Tomando en cuenta la comparación de estructuras tridimensionales construyeron la base de datos Clasificación Estructural de las Proteínas (SCOP, por sus siglas en inglés) mientras que Thornton *et al*[4]. produjeron la base de datos Clase(C), Arquitectura(A), Topología(T) y superfamilia homóloga (h) (CATH, por sus siglas en

inglés). Ambas bases de datos pretenden ser una clasificación integral, es decir, toman en cuenta la estructura, la función y el porcentaje de similitud en la secuencia. Los niveles de jerarquía están definidos por la estructura tridimensional, mientras que los taxa inferiores están identificados con base en la similitud de la secuencia y en consideraciones funcionales. Aun cuando en la década de 1960, el grupo de Margaret Dayhoff[5], introdujo la noción de familia y súper familia, definiendo a la familia como un grupo de proteínas que comparten un posible ancestro común, mientras que la súper familia engloba a dos o más familias relacionadas estructuralmente[6], Murzin[7], define una familia como aquel grupo de proteínas que comparte al menos 30% de identidad en su estructura primaria, o bien, como grupos de proteínas con baja identidad en la secuencia pero cuyas funciones y estructuras son muy semejantes; por ejemplo, las globinas con identidades en la secuencia del 15%, mientras que a la súper familia la define como familias cuyas proteínas tienen baja identidad en la secuencia pero cuyas estructuras y en muchos casos, las características funcionales apoyan que un origen evolutivo común es posible.

1.2 Relación estructura-función

Como sabemos, las proteínas están constituidas por cadenas largas, en las cuales los aminoácidos se encuentran formando secuencias lineales específicas. En cada tipo de proteína la cadena polipeptídica se halla plegada adoptando una conformación tridimensional específica, la cual es necesaria para su función biológica o actividad. Las moléculas de proteína sólo poseen habitualmente una conformación tridimensional específica en las condiciones normales intracelulares, es decir, en el citoplasma, denominada conformación nativa. Se puede determinar esta conformación mediante el análisis por difracción de rayos X o por resonancia magnética nuclear.

El espaciado de unidades moleculares o atómicas, repetidos regularmente en cristales, puede determinarse estudiando los ángulos y las intensidades a que los rayos X de una longitud de onda determinada son dispersados o difractados por los electrones que rodean a cada átomo. Los átomos que poseen las densidades electrónicas más elevadas, tales como los metales pesados, producen la máxima difracción de los rayos X y los átomos de hidrógeno, que poseen la menor densidad electrónica, son los que difractan menos los rayos X.

El análisis por rayos X de cierto número de proteínas globulares muestra que cada tipo de proteína tiene un modo particular de plegarse, es decir, una estructura terciaria diferente. La estructura tridimensional de las proteínas contiene dominios, unidades estructurales mínimas de construcción, que provienen de un mismo ancestro[8] y pueden ser unificados en un grupo de plegamientos. Por lo cual, el universo de la topología proteica es finito y estructuralmente redundante, siendo los plegamientos existentes los elementos más conservados en la biología. El gran interés de la bioquímica y de la biología molecular consiste en descifrar la relación que priva entre la estructura tridimensional de las proteínas y su función biológica.

Ahora bien, la mayoría de las proteínas que exhiben similitudes estructurales significativas, llevan a cabo funciones idénticas o similares. Más allá de dichas similitudes inherentes, las diferentes funciones enzimáticos (definidas por sus números asignados por la Comisión Enzimática, EC, por sus siglas en inglés) son llevadas a cabo por proteínas que poseen una amplia variedad de arquitecturas y topologías diferentes.

Hasta 1998, todos los pares de proteínas con secuencias que indican una relación evolutiva evidente, adoptan el mismo plegamiento, solo con pequeñas variaciones (Ej. cambios en la orientación de los dominios, largo de las asas o estructuras secundarias adicionales). Por ejemplo, las globinas de varias especies con secuencias ampliamente divergentes, adoptan el mismo plegamiento y llevan a cabo la función de transporte de oxígeno.

Sin embargo, también hay proteínas homólogas que tienen claramente funciones diferentes, a pesar de que adoptan la misma estructura. El ejemplo clásico es de la lisozima y la α-lactalbúmina. Aunque estas enzimas poseen ~35% de identidad en la secuencia, la α-lactalbúmina ha perdido el carboxilo catalítico del cual los residuos de glutamato y aspartato son necesarios para la ruptura del azúcar.

En contraste, hay varios ejemplos de proteínas que llevan a cabo la misma función, pero es claro que no están relacionadas evolutivamente. Aquí, el ejemplo clásico lo constituyen la tripsina y la subtilisina, las cuales no solo llevan a cabo la misma función, a pesar de tener estructuras totalmente diferentes, sino que han desarrollado el mismo mecanismo con la

misma tríada catalítica Asp-His-Ser. Este es un claro ejemplo de convergencia funcional. Un ejemplo diferente es provisto por los inhibidores de proteasa dependientes de serina. Aunque estas proteínas adoptan una amplia variedad de plegamientos, todas poseen una "estructura de asa" que mimetiza al sustrato y se une de manera covalente al sitio activo de la proteinaza. También hay proteínas con función dual, quizá el ejemplo más extremo es el de la tripsina en el virus Sinbdis, el cual realizando su función catalítica se convierte en la proteína de la capside del virus. Este ejemplo sirve para enfatizar que las relaciones entre la estructura y función no son directas y un análisis global solo revela ligeras tendencias a las cuales siempre hay excepciones.

En un estudio realizado en 1998[9], demuestran que hay una mínima correlación entre la clase de proteína o arquitectura y la función enzimática, presumiblemente por que la actividad enzimática está definida sólo por unos pocos aminoácidos. En contraste, parece haber mucho mejor correlación entre la clase de arquitectura y el tipo de ligando. Esto lo determinan, en las enzimas, dado que son la clase funcional de proteínas más fácil de analizar en el PDB por que son numerosas (5819 cadenas en julio de 1997 y también asignadas en el CATH) y están clasificadas en términos funcionales por sus números EC. En el trabajo solo se consideraron la clase primaria para una cadena enzimática de un solo dominio (lo cual evita el problema de asignar la actividad enzimática a un dominio específico). En las enzimas, la actividad catalítica y función dependen de la localización y orientación específica de unos pocos aminoácidos. Por lo tanto, de todas las proteínas, las enzimas son las que menos relaciones fundamentales exhiben entre su estructura (completa) embebidas en los niveles más altos del CATH y su función específica. De hecho, entre las enzimas de un solo dominio en la base de datos se encontraron 37 ejemplos de un número EC correspondiendo a más de una topología (incluyendo 5 ejemplos del mismo número EC siendo asignado a 4 plegamientos diferentes) y 36 ejemplos de miembros de una sola súper familia homóloga con diferentes números EC.

1.3 El barril TIM

Hasta 1990[10], se pensaba que enzimas estructuralmente relacionadas, catalizaban reacciones químicas idénticas, pero con sustratos diferentes, es decir prácticamente un plegamiento para cada tipo de reacción catalítica, sin embargo, en ese año, se vió que las

enzimas mandelato racemasa y la muconato lactonizante adoptan el mismo plegamiento, el de tim barrel y catalizan reacciones diferentes, lo cual atrajo inmediatamente la atención hacia este plegamiento.

Existen plegamientos que realizan muy variadas funciones, como es el caso del barril TIM, las actividades enzimáticas que realizan, incluyen cinco de las seis clases primarias definidas por la Comisión Enzimática (EC)[11]. Esto nos indica lo exitoso que resulta este plegamiento, dado que prácticamente puede catalizar cualquier tipo de reacción enzimática. La única función que no se ha encontrado aún es la de ligaza.

Este plegamiento es uno de los más frecuentes y regulares de los dominios β/α, consiste de una estructura central formada por hojas beta paralelas o mixtas (antiparalelas) rodeadas por hélices alfa y unidas entre sí mediante asas, las cuales si bien no contribuyen a la estabilidad estructural, participan en la unión al sustrato y a la actividad catalítica. Recibe el nombre de barril TIM debido a la triosa fosfato isomersa, la enzima en la cual fue observado por primera vez[12]. La figura 1. Muestra una representación gráfica de la triosa fosfato isomersa, un barril (c_8 típico, mientras que en la tabla 1, se muestran las funciones enzimáticas que realizan todas las enzimas que adoptan este plegamiento.

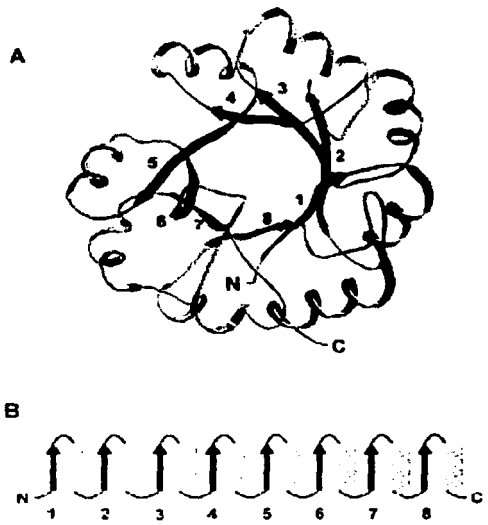


Figura 1. Vista esquemática de la Triosa fosfato isomersa de *Gallus gallus*. (A) Se anota la localización de los C-terminales de las ocho hojas beta, las cuales están rodeadas por ocho hélices alfa. (B) Topología de los ocho módulos (β/α). El sitio activo de todas las enzimas conocidas de barril TIM se localiza en el C-terminal de las hojas beta, se ilustran los "loops" que conectan a cada hoja beta con la subsecuente hélice alfa.

Tabla 1. Funciones realizadas por las enzimas que adoptan el plegamiento barril TIM, acotadas por su número E.C.

Súper familia SCOP	Familia	Número de E.C.
1. Triosa fosfato isomersa.	Triosa fosfato isomersa.	5.3.1.1
2. Enzimas que unen ribulosa fosfato	Enzimas de la biosíntesis de His	3.1.3.15
	D-ribulosa-5-fosfato 3-epimerasa	5.1.3.1
	Descarboxilasa	4.1.1.23
	Enzimas de la biosíntesis de Trp	4.1.1.48
3. Tiamina fosfato sintasa	Tiamina fosfato sintasa	2.5.1.3
4. Piridoxin 5' fosfato	Piridoxin 5' fosfato	2.7.1.35
5. Óxidorreductasa que unen FMN	Óxidorreductasa que unen FMN	1.5.99.7
6. Inosin monofosfato deshidrogenasa	Inosin monofosfato deshidrogenasa	1.1.1.205
7. Enzimas que unen PLP	Dominio, N-terminal semejante a la alnina racemasa	5.1.1.1
	Proteína hipotética yb1036c	
8. Óxidoreductasas que unen NADP	Aldo-ceto reductasas (NADP)	1.1.1.21
9. Transglicosidasas	Amilasa	3.2.1.1
	Beta-glucanasas	3.2.1.39
	Glicosilhidrolasa	3.2.3.1
	Quitinasas tipo II	3.2.1.14
	1,4-beta-N-acetilmuraminidasa	3.2.1.17
	Beta-N-acetilhexosaminidasa	3.2.1.52
	Alfa D-glucoronidasa	3.2.1.139
	Beta D-glucano hexohidrolasa	3.2.1.58
	Bee venom hialuronidasa	3.2.1.35
10. Hidrolizas metal-dependiente	Anosina desaminasa	3.5.4.4
	Dihidroorotasa	3.5.2.3
	Citosina desaminasa	3.5.4.1
	Subunidad alfa de la ureasa	3.5.1.5
	Hidantoinasa	3.5.2.2
	Proteína hipotética TM0936	
	N-acetilglucosamina-6-fosfato desacetilasa	3.5.1.25
	D-aminocilasa	
	TatD dependiente de Mg parecida a una Dnasa	3.5.1.81
	Fosfotriesterasa	
	Dipeptidasa Renal	
	Uronato isomerasa	3.1.8.1
		3.4.13.19
		5.3.1.12
11. Aldolasa	Clase I	4.1.2.4
	Clase II FBP	4.1.2.13
	5-aminolevulinato deshidratasa	4.2.1.24
	Clase I DAHP sintasa	4.1.2.15
12. Dominio enolasa C-terminal	Enolasa	4.2.1.11
	D-glucarato deshidratasa	4.2.1.40
13. Fosfoenolpiruvato	Piruvato cinasa	2.7.1.40
	Piruvato fosfato dicinasa	2.7.9.1

	Fosfoenopiruvato carboxilasa	4.1.131
	Fosfoenopiruvato mutasa	5.4.2.9
	2-deshidro-3-desoxi-galactarato aldolasa	4.1.2.20
	Isocitrato liasa	4.1.3.1
14. Malato G sintasa	Malato G sintasa	4.1.3.2
15. Rubisco	Rubisco	4.1.1.39
16. Xilosa isomersa	Endonucleasa IV	3.1.21.2
	Proteína hipotética Ioll	
	Proteína hipotética YgbM	
	L-ramnosa isomerasa	5.3.1.14
	Xilosa isomerasa	5.3.1.5
17. Luciferaza semejante a la de bacterias	Luciferaza bacteriana	1.14.14.3
	Flavoproteína no fluorescente	
	Tetrahidrometanopterina reductasa dependiente de Coenzima F420	
	Alkanesulfonato monoxigenasa SsuD	1.1.-.-
18. Ácido quinolínico PR-transferasa	Acido quinolínico PR-transferasa	2.4.2.19
19. Fosfolipasa fosfatidilinositol	De mamíferos	3.1.4.11
	De bacteria	3.1.4.10
20. Enzimas dependientes de cobamlamina vitamina B12	Metilmalonil CoA mutasa	5.4.99.2
	Glutamato mutasa	5.4.99.1
	Diol-deshidratasa	4.2.1.28
21. tRNA-guanín transglicosidasa	tRNA-guanín transglicosidasa	2.4.2..29
22. Dihidropteroato sintasa	Dihidropteroato sintasa	2.4.2.29
23. Uroporfirinogeno dsecarboxilasa	Uroporfirinogeno dsecarboxilasa	4.1.1.37
24.Metilenetetrahidrofolato reductasa	Metilenetetrahidrofolato reductasa	1.7.99.5
	Prolina deshidrogenasa	1.5.99.8
25. Monometilamin metiltransferase	Monometilamin metiltransferase	2.1.1.-
26. Betain-homocisteina S-metiltransferasa	Betain-homocisteina S-metiltransferasa	2.1.1.5

A medida que más estructuras tridimensionales han sido resueltas, se ha observado que el mismo motivo ha sido empleado una y otra vez para una variedad de funciones. Así, pues, al ver la amplia gama de reacciones catalíticas que realiza este plegamiento y lo difundido que se encuentra en el espacio proteico, Farber y sus colaboradores[13] en 1990 se preguntan acerca de las relaciones evolutivas de estas enzimas. La disyuntiva que plantean es si este grupo de enzimas están relacionadas por evolución divergente a partir de un ancestro común o por evolución convergente a partir de varios ancestros que poseen plegamientos β/α semejantes.

Para resolver la pregunta planteada, Farber y sus colaboradores [12], realizan un análisis de las 17 enzimas, reportadas a esa fecha, que poseían un dominio del tipo barril $(\beta/\alpha)_8$, en el

cual, concluyen que pueden ser agrupadas en cuatro familias estructurales y que son producto de una evolución divergente al interior de cada una de ellas, pero no logran establecer un ancestro evidente entre ellas. Uno de los argumentos más sólidos que presentan en favor de la evolución divergente a partir de un ancestro común, es el hecho de que todas las proteínas conocidas que adoptan un plegamiento barril $(\beta/\alpha)_8$ son enzimas. El segundo argumento, es la localización del sitio activo de esas enzimas. A pesar de la gran diversidad de reacciones químicas que catalizan, el sitio activo siempre está localizado en el C-terminal de la hoja β-barril. Finalmente, lo que ellos logran encontrar es que pueden agrupar las 17 enzimas en cuatro familias estructurales pero lejanas entre sí, quedando pues la pregunta de si será posible distinguir entre cuatro familias primordiales de barril $(\beta/\alpha)_8$ y un ancestro común a todas ellas.

1.4 Mutaciones

Los eventos primarios en la generación de la diversidad biológica son la mutación, la inserción y la deleción de nucleótidos en la secuencia del DNA o la transposición, a gran escala, de alguna pieza en el material genético y que la selección reacciona a la función de la proteína según lo determinado por su misma estructura.

Con respecto a la selección, las mutaciones pueden ser neutras, deletéreas y selectivas a favor o en contra. Cuando ocurre un cambio en la secuencia de aminoácidos de una proteína, puede ser que el cambio se dé por un aminoácido del mismo tipo (considerando la siguiente clasificación de los aminoácidos: neutros o alifáticos, aromáticos, hidroxiaminoácidos, tioaminoácidos, iminoácidos, dicarboxílicos y sus amidas y bibásicos) o bien por uno que sea de naturaleza diferente. Dependiendo del tipo de cambio que ocurra, éste puede ser benéfico o deletéreo, en función de ser afectada la flexibilidad de la proteína o el sitio activo de la misma.

En la década de 1960, se realizaron algunas observaciones, tales como la existencia de una divergencia molecular proporcional al tiempo de separación de las especies, que la tasa de evolución guarda una relación inversa con la importancia funcional del gen, región o proteína y finalmente, los altos niveles de polimorfismo revelados por electroforesis de proteínas, que llevaron a formular la teoría neutral de la evolución molecular por

Kimura[14,15], donde se postulaba que las sustituciones que se fijan en la evolución y los polimorfismos alélicos que se encuentran en las poblaciones, son neutras.

Una nueva variante surgida por mutación se encuentra inicialmente en un único individuo de la especie. Con el transcurso del tiempo esta variante puede pasar a tener una frecuencia del 0 (pérdida) al 100% (fijación) en la especie estudiada, y en algunos casos puede permanecer en la población con una frecuencia intermedia (polimorfismo). Así, pues, son las mutaciones que permanecen en la población ya sean las que se han fijado (sustituciones) o las que se encuentran en las frecuencias intermedias las que provocan las diferencias detectadas[12,13]. De esta manera, al paso del tiempo, un par de proteínas cuya secuencia de aminoácidos originalmente era muy parecida, al fijarse las mutaciones e irse acumulando, pueden diferir tanto en la secuencia, que el parecido inicial puede llegar a niveles donde apenas se pueda reconocer la homología. Se dice entonces que han divergido. El grado de divergencia con una mutación puede rebasar el umbral en donde se puede detectar la homología, lo cual no necesariamente implica que las proteínas no sean homólogas.

1.5 Bioinformática

Debemos empezar por definir a la bioinformática, de la siguiente manera: un campo interdisciplinario que involucra a la biología, las ciencias de la computación, las matemáticas y la estadística para analizar los datos de las secuencias biológicas, el contenido de los genomas y su arreglo y predecir la función y estructura de las macromoléculas. El mayor motor para el desarrollo de la bioinformática ha sido la biología molecular y en concreto, la posibilidad de disponer de secuencias de DNA y de proteínas. Las herramientas bioinformáticas más usadas han sido desarrolladas como una respuesta a las necesidades de obtener información sobre las secuencias aprovechando el conocimiento previo almacenado en bases de datos.

Habitualmente se dispone de cierto conocimiento sobre la secuencia problema con lo que se puede extraer de las bases de datos información relacionada con la secuencia problema bien buscando otras secuencias similares a ella o bien, buscando secuencias cuya información asociada tenga algo en común con la de la secuencia problema. Alternativamente se pueden buscar motivos funcionales, esto es, pequeñas regiones, caracterizadas previamente, con un significado funcional o estructural en la secuencia problema. Utilizando estos motivos se pueden buscar en la base de datos nuevas

secuencias, no detectadas previamente en los pasos anteriores, y que estén relacionadas funcionalmente o estructuralmente con la secuencia problema. Con la información obtenida de las bases de datos se puede dar un paso más y realizar un alineamiento múltiple. La definición de zonas conservadas y variables, junto con el conocimiento de las propiedades de las distintas secuencias en el alineamiento permiten, en muchos casos, definir que regiones están implicadas en la función de las proteínas. Igualmente, la derivación de una árbol genealógico relacionando secuencias distintas pero emparentadas puede dar mucha información acerca de las zonas donde reside la funcionalidad de las secuencias. También se puede definir un motivo característico del alineamiento obtenido y usarlo para recuperar secuencias, que a pesar de haber perdido la similitud a nivel global, contengan esta región. Así, pues, al encontrar homología en las secuencias, no solo nos damos cuenta de que obtenemos información acerca de la función que realizan, sino también del plegamiento que adoptan.

Con base en lo anterior, resulta evidente que la búsqueda de secuencias de proteínas en una base de datos es un método muy eficiente para obtener información sobre la estructura y la función de estas moléculas.

1.6 Algoritmos de búsqueda

La dificultad principal con la búsqueda es determinar la significancia de los alineamientos que son encontrados. Tales similitudes pueden ser evidencia de relaciones estructurales o evolutivas, pero también pueden ser apareamientos de variaciones azarosas que no tienen un origen común o función. Por tal razón, la similitud en la secuencia no es, usualmente, un buen indicador de similitud estructural y los alineamientos encontrados deben ser evaluados cuidadosamente antes de afirmar cualquier conclusión.

Un método muy empleado para determinar las distancias que existen entre secuencias es la matriz PAM250 (Mutaciones puntuales aceptadas, por sus siglas en inglés) el cual está basado en la estimación de frecuencia de transición para una pequeña cantidad de cambio en la secuencia (típicamente 1%). Una distancia evolutiva de 1 PAM indica la probabilidad de que un residuo mute en una distancia en la cual una mutación puntual fue aceptada por 100 residuos. Las matrices PAM25, proveen un mejor alineamiento para proteínas lejanamente relacionadas del 14 al 27% y están así en la zona desconocida (dimensión

desconocida). Por lo cual, en estos casos, se recurre a una comparación de las estructuras tridimensionales con ayuda de uno de los varios métodos empleados para tal fin, el de extensión combinatoria (CE, por sus siglas en inglés) dicho programa, alinea dos cadenas del polipéptido usando características de su geometría local según lo definido por vectores entre las posiciones de los carbonos alfa. Los grupos fosfato se llaman los pares alineados del fragmento (AFPs). La heurística se utiliza en definir un sistema de trayectorias óptimas que ensamblan AFPs con gaps según sea necesario. La trayectoria con la menor RMSD, (root mean square deviation) está conforme a la programación dinámica para alcanzar una alineación óptima. Así, cuando tenemos el caso de un par de proteínas que tienen un bajo porcentaje de identidad en la secuencia, pero un buen RMSD, es decir menor de 3.0 Å, podemos decir con certeza que son claramente homólogas.

Existe una gran variedad de algoritmos para realizar búsquedas basados en alinear la secuencia problema con todas las secuencias contenidas en la base de datos de forma que se pueden extraer aquellas que muestran un parecido que se refleje en dicho alineamiento. Los algoritmos tradicionales garantizan que encuentran el alineamiento óptimo, basándose, principalmente, en métodos de programación dinámica[16,17]. Sin embargo, estos métodos, requieren de un tiempo $O(N^2)$ es decir, el tiempo calculado es proporcional al cuadrado del tamaño de las secuencias que se comparan. En las bases de datos se suele rastrear buscando un parecido local, ya que es probable que los homólogos difieran a nivel global (debido a nuevos arreglos). Conforme crecieron las bases de datos de secuencias, se volvió impráctico rastrearlas en forma exhaustiva mediante programación dinámica[12]. Por ello se desarrollaron métodos no-exhaustivos, llamados heurísticos, que buscan velocidad a cambio de sensibilidad y selectividad. Éstos, hacen una preselección rápida de posibles homólogos, eliminando a la gran mayoría de las secuencias de la base de datos. Las pocas elegidas pasan a una evaluación estricta, que a menudo incluye programación dinámica. Para reducir el espacio de búsqueda (y por tanto el tiempo de computación) se intenta localizar las posiciones donde es más probable que se encuentre el mejor alineamiento (posición de mayor coincidencia entre las secuencias). Por ejemplo, la serie de programas FASTA[18] basa su estrategia en identificar diagonales con el mayor número de identidades que luego analiza usando el esquema de puntuación.

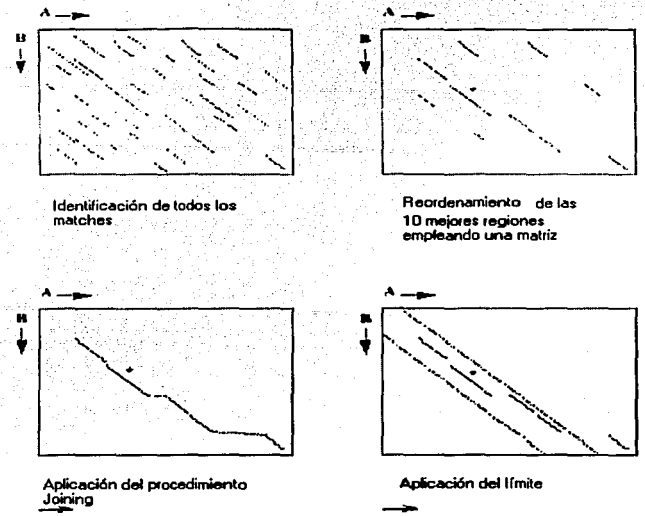
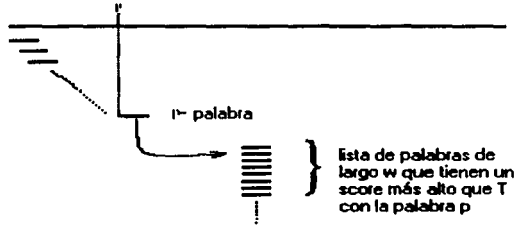
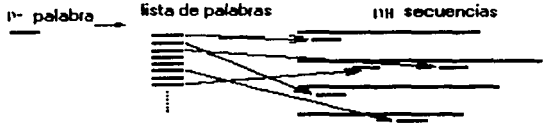


Figura 2. Esquema del proceso realizado por FASTA.

A: Para cada posición p de la semilla encuentra la lista de palabras de largo w con un score más alto que T y entonces lo aparea con la palabra iniciando en p



B: Para cada una de las palabras de la lista identifica los "matches" más altos con las secuencias DB



C: Para cada palabra con un "match" ("hit") extiende un alineamiento sin "gaps" en ambas direcciones. Parando entonces las disminuciones en S, por más que X, a partir de los valores más altos, haya sido alcanzado por S

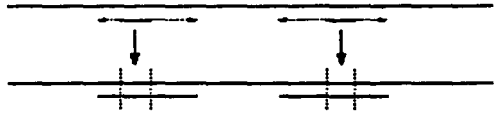


Figura 3. Esquema del funcionamiento de BLAST.

De la misma forma que FASTA busca fragmentos de k residuos que son idénticos con las secuencias comparadas, para identificar diagonales potencialmente interesantes (perdiendo información de los fragmentos no idénticos o más pequeños que k), BLAST amplía la búsqueda a todos aquellos fragmentos que producen emparejamientos positivos. Puesto que esto amplía el espacio de búsqueda, una forma de volverlo a reducir es descartando aquellos emparejamientos que estadísticamente tengan una probabilidad baja de ocurrir. Para ello es necesario tener en cuenta dos aspectos básicos: en primer lugar, la identificación rápida de segmentos con una puntuación alta (no necesariamente idénticos) y por otra parte, la elección (para su análisis en detalle) de aquellos segmentos que tengan una probabilidad más alta de estar contenidos en el alineamiento final, descartando aquellas puntuaciones que pueden ser debidas al azar[19]. Ambos métodos de búsqueda emplean una valoración estadística, siendo ésta su gran contribución.

1.7 Búsquedas empleando perfiles como semillas

Numerosos estudios han demostrado que las búsquedas en las bases de datos empleando matrices de peso posición-específica (PSSMs, por sus siglas en inglés) o perfiles como problemas (o semillas) son más efectivos en la identificación de relaciones distantes entre proteínas que aquéllas búsquedas que usan solo una secuencia como problema.

Existen ventajas en emplear una matriz de calificación que representa los patrones conservados de la secuencia en una familia de proteínas, en lugar de una sola secuencia problema para buscar en una base de datos. Por lo tanto, la búsqueda en la base de datos, puede extenderse para identificar secuencias relacionadas que de otra manera se perderían. La mayor dificultad con este tipo de búsquedas es que un alineamiento de secuencias relacionadas debe estar disponible para conocer las variaciones en cada una de las posiciones en la secuencia problema. Una nueva versión de BLAST, denominada Posición-Específica Iterado BLAST (**PSI-BLAST**) ha sido diseñado para proveer información sobre esta variación, iniciando con una búsqueda de BLAST para una sola secuencia problema. El método usado por PSI-BLAST involucra una serie de pasos repetidos o iterados. Primero, se realiza una búsqueda en la base de datos partiendo de una secuencia problema. Segundo, el resultado de la búsqueda es presentado y pueden ser valorados visualmente entre pocas secuencias que están relacionadas significativamente a nuestra secuencia problema. Tercero, si es el caso, se lleva a cabo otra iteración. Los registros altos en las

secuencias apareadas encontradas en el primer paso son alineados y a partir del alineamiento, un tipo de matriz de calificación (PSSM) que indica las variaciones en cada una de las posiciones alineadas es producida. Se busca nuevamente en la base de datos con esta matriz. Así, la búsqueda ha sido extendida para incluir secuencias que aparean las variaciones encontradas en los alineamientos múltiples a cada una de las posiciones. Los resultados se muestran nuevamente, indicando las secuencias nuevas que están relacionadas significativamente a las secuencias alineadas en la iteración previa. Se puede llevar a cabo otra iteración del programa, incluyendo nuevas secuencias reclutadas. De esta manera, una nueva familia de secuencias son significativamente semejantes a la secuencias problema, pueden ser encontradas.

Con ayuda de este programa, podemos localizar homólogos que con otros métodos no podríamos encontrar. Si recordamos lo que habíamos definido sobre el concepto de familias de proteínas, el parecido entre las proteínas de una familia y otra es menor que entre las proteínas de una misma familia, siendo estos valores de parecido muy variables según la familia de que se trate. Teniendo en cuenta esto podemos pensar que el "universo proteico" está formado por "islas" (que serían las familias) y la distancia entre éstas depende de si tienen una relación evolutiva o no y si la tienen, de cuán parecidas son. Decimos que hay una relación de homología si existe parecido alguno que demuestre que hay un origen evolutivo común. En este contexto, debemos revisar la idea de que la homología es transitiva. Dos secuencias homólogas, las cuales han divergido más allá del punto donde su homología puede ser detectada mediante una simple comparación directa, pueden ser relacionadas mediante una tercer secuencia que se sitúe entre ambas. La presencia de registros altos entre la secuencia intermediaria y la primera, así como entre la intermediaria con la tercera, implica que la primer y tercer secuencias, están relacionadas a pesar de que los registros entre ellas sean bajos[20], siempre y cuando sea en la misma región.

2. Antecedentes

Copley y colaboradores[21] empleando la base de datos SCOP y ASTRAL[22] seleccionaron secuencias representativas de barril TIM, al nivel de 50% de identidad, puesto que de algunas, no tenían la estructura tridimensional y así aseguraban que eran homólogas. Estas secuencias fueron empleadas para iniciar una búsqueda en la base de datos no redundante (NR) del NCBI, con el programa PSI-Blast[23], con un valor de esperanza inclusión <0.001,

que es que el tiene determinado el programa, encontrando una clara relación entre 12 súper familias de 24. Las familias se denotan con una C en la Tabla 2 de la página 17. Realizaron un alineamiento de las hojas 6 y 7 para varios barriles TIM, empleando el paquete STAMP[24]. Encuentran secuencias con un residuo de lisina conservada en la hoja 6, de la familia de las aldolasas.

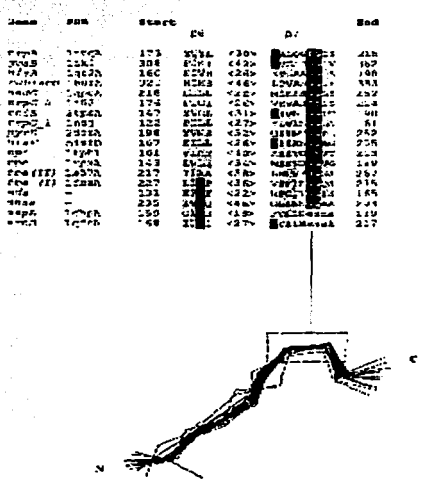


Figura 4. Alineamiento que realizaraon Copley y colaboradores[21].

En un estudio posterior, Nagano y colaboradores[25] tomaron como referencia la base de datos CATH, y encontraron una relación entre 21 de las 26 súper familias que maneja dicha base de datos. Emplearon el programa PSI-Blast para detectar relaciones sutiles entre las súper familias, con un número máximo de 20 iteraciones, con un valor de E <0.005 y lo hacen contra una base de datos de las secuencias de aminoácidos de las proteínas contenidas en el PDB. El programa IMPALA[26] es una versión más refinada del PSI-Blast y provee alineamientos mejores que el PSI-Blast original, usándolo también para obtener alineamientos a través del Gen Bank. En la Tabla 2 de la página 17, se resaltan con una N las familias para las cuales detectan homología. Realizan también alineamientos usando la técnica CORA[27], la cual genera alineamientos múltiples de las estructuras tridimensionales. Con base en lo anterior, ellos logran establecer varias secuencias intermedias entre un grupo bien definido de ciertas súper familias, las (tras)glicosidasas, las quitobiasas, las oxidorreductasas dependientes de NADP y las fosfolipasas C. Las restantes diecisiete resultan homólogas en torno a este grupo. Sin embargo, existen dos súper familias cuyo agrupamiento no es tan sólido, pues son base en un par de residuos conservados,

argumentan que podrían unirse al grupo antes mencionado: la súper familia de la xilosa isomersa y las hidrolasas metal-dependientes.

Ambos trabajos presentan una limitación en sus resultados, derivada de la base de datos empleada para realizar la búsqueda, las bases de datos SCOP, ASTRAL y pdb no son muy grandes, pues son de estructuras, por esta razón se decidió realizar una búsqueda empleando la base de datos No Redundante.

Así pues, el presente trabajo trata de entender que tipo relación evolutiva existe entre las diferentes súper familias de proteínas que adoptan el plegamiento barril TIM.

3. Hipótesis

Las proteínas con estructura semejante son producto de evolución divergente, por lo cual esperamos que todas las proteínas que adoptan un plegamiento del tipo barril TIM, resulten ser homólogas.

4. Objetivo

Determinar si las súper familias que adoptan el plegamiento barril TIM, para las cuales no se ha detectado una homología evidente, se relacionan evolutivamente.

5. Método

Se usó la base de datos SCOP[3] tomando todas las secuencias disponibles para las proteínas clasificadas en las diversas súper familias de baril TIM para realizar una búsqueda en la base de datos NR del NCBI empleando el programa PSI-BLAST [8]. Se escogió este programa debido a la rapidez con la que se pueden realizar los análisis. El número máximo de iteraciones fue de 20, con un umbral de inclusión <0.001. Se hace necesario analizar cada salida de los PSI-Blast, para lo cual se realizan un par de "scripts". En el apéndice se muestran los scripts empleados.

En la figura 4 se resumen la serie de pasos que se siguieron para realizar las búsquedas.

Una vez que sabemos que existe homología entre cada par de enzimas, con estos archivos generamos una matriz que nos indica el número de secuencias intermedias encontradas para cada par de secuencias analizadas. En la Tabla 2 se muestran las enzimas empleadas para el análisis del plegamiento barril TIM.

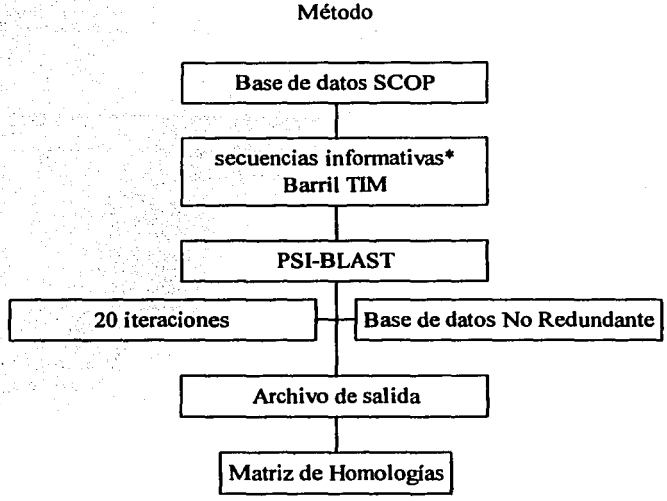


Figura 4. Diagrama de flujo ilustrando el método empleado.

Tabla 2. Lista de proteínas y códigos del PDB empleados en el presente trabajo. El número entre paréntesis indica cuántas familias poseen.

Súper familia SCOP	Abreviatura	RN	No. Fam.	No. sec.	No. sec. empleadas	Entrada PDB
1. Triosa fosfato isomersa.	TIM	C N	1	15	2	1tph, 1tpfA
2. Enzimas que unen ribulosa fosfato	ERF	C N	4	19	5	1rpx, 1nsj, 1a53, 1ubsA, 1ttq
3. Tiamina fosfato sintasa	TFS	C N	1	8	1	2tps
4. Piridoxin 5' fosfato	P5F	*	1	6	1	1hol
5. Oxidorreductasa que unen FMN	OFMN	C N	1	16	11	2dor, 1ltd 1fcbA, 1gox, 2tmdA, 1oya, 1dorA, 1rpxA, 1pii, 1a53, 1nsj
6. Inosin monofosfato deshidrogenasa	IMD	N	1	5	1	1ak5
7. Enzimas que unen PLP	EPLP	N	2	5	1	1bd0A
8. Oxidoreductasas que unen NADP	ONADP	C N	1	10	3	1lwi, 2alr, 1ads
9. Transglicosidasas	TG	N	9	82	34	1ava, 2aaa, 1ppi, 1bli, 1bag, 1amg, 1bf2, 1bvzA, 1cgt, 1uok, 1byb, 1b9z, 1ceo, 1eceA,

						1edg, 1egzA, 1xyzA, 1gow, 1bglA, 1bhg, 1bqcA, 1pbgA, 2xyl, 1ghs, 1aq0A, 1cbg, 2myr, 1ctn, 2hvm, 2ebn, 1nar, 1qba, 1tml, 1cb2.
10. Hidrolizas metal-dependiente	HMD	C N	8	18	5	pyrC 1a4mA, 2kauC, 1pscA, 1bf6A
11. Aldolasa	AL	C N	4	29	6	1fba, 1b57, 1onr, 1qr7, 1dhp,1nal
12. Dominio enolasa C-terminal	DEC	C N	2	13	5	1one, 1muc, 1mdl, 1chrA, 1bqg
13. Fosfoenolpiruvato	PEP	C N	6	14	4	1a3w, 1dik, 1fiy, 1pkm
14. Malato G sintasa	MGS	C N	1	2	1	1d8c
15. Rubisco	RUB	N	1	8	1	1rbl
16. Xilosa isomersa	XI	N	5	16	2	1a0dA, 1xib
17. Luciferaza semejante a la de bacterias	LB	N	3	6	3	1fvpA, 1lucA, 1lucB
18. Ácido quinolínico PR-transferasa	AQPRT	C N	1	2	1	1qpo
19. Fosfolipasa fosfatidilinositol	FF	N	2	3	3	1gym, 1qasA, 2plc
20. Enzimas dependientes de cobamlamina vitamina B12	EC	N	3	4	1	1qtwA
21. tRNA-guanín transglicosidasa	tRNAGT	N	1	2	1	1wkf
22. Dihidropteroato sintasa	DS	C N	2			1ajz 1ad4B, 1f6y
23. Uroporfirinogeno dsecarboxilasa	UD	*	1			1uro
24.Metilenetetrahidrofolato reductasa	MR	*	1			1b5t
□25. Monometilamin metiltransferase	MM	*	1			1l2r
□26. Betain-homocisteina S-metiltransferasa	BHM	*	1			1lt8

Únicamente se tomaron las proteínas informativas. En algunos casos se muestra la cadena empleada en este trabajo. ℵR=Referencia, la C denota a las súper familias para las cuales Copley y colaboradores detectan homología, mientras que la N, marca a las súper familias reportadas por Nagano *et al.* *súper familias para las que no se había reportado homología □ Súper familias recientemente reportadas.

6. Resultados y discusión

6.1 Súper familias para las que no se había detectado homología

La súper familia de la metilenetetrahidrofolato reductasa es identificada como homóloga de las súper familias aldolasas y enzimas que unen ribulosa fosfato, mediante el análisis realizado con PSI-BLAST, el cual produjo valores de *E* significativos a partir de la búsqueda iniciada con la secuencia de la metilenetetrahidrofolato reductasa de *E. coli*. Esta enzima, la N^5, N^{10}-metilenetetrahidrofolato reductasa, cataliza la reducción irreversible del N^5, N^{10}-metileneH_4folato hacia 5-metil derivado, reacción necesaria en la síntesis de metionina a partir de homocisteína. Estudios previos[7,28] demuestran que la homología entre los miembros de la súper familia de enzimas que unen ribulosa fosfato, concretamente entre las enzimas de la síntesis de histidina y las enzimas de síntesis de triptofano. Por último, la nueva súper familia, Betain-homocisteina S-metiltransferasa, se detecta como homóloga tanto de la súper familia de las enzimas que unen cobalamina y de la metilenetetrahidrofolato reductora. Así, pues, se detecta una homología entre cuatro súper familias involucradas en la síntesis de aminoácidos.

Entre la súper familia uroporfirinogeno descarboxilasa y las enzimas dependientes de cobalamina se detecta también homología.

La súper familia piridoxin 5'-fosfato sintasa se detecta como homóloga de la súper familia tiamina fosfato sintasa a partir del análisis realizado con PSI-BLAST, la enzima PdxK de *E. coli* P40191 fosforila los esqueletos de vitamina B6, mientras que la ThiE de *Bacillus subtilis* cataliza la reacción: 2-metil-4-amino-5-hidroximetilpirimidin di fosfato + 4-4-metil-5-2-fosfonooxietil-tiazole = di fosfato + tiamina monofosfato. Ambas enzimas requieren de la unión a un cofactor, el cual puede ser zinc o magnesio en el caso de la PdxK de *E. coli*, mientras que la ThiE de *B. subtilis* liga un ión de magnesio por subunidad.

En la tabla 4, de la página 23, se presentan las súper familias de barril TIM, resaltando con una C aquellas homologías detectadas por Copley *et al.* y en N las detectadas por Nagano *et al.*, finalmente, con una L aquellas relaciones establecidas en el presente análisis. La Figura 2, provee una visión de las relaciones entre las súper familias encontradas a través de las

secuencias seleccionadas, listadas previamente en la Tabla 2. Cada una de las súper familias tiene un número y una abreviatura, empleados a lo largo de este escrito. Doce de las 26 súper familias consisten de sólo una secuencia. La súper familia más grande es la 9. Transglicosidasas, que tiene treinta y tres secuencias. La siguiente súper familia es la 5, Óxidorreductasas que unen FMN, con doce secuencias. Se toman en cuenta solo las proteínas informativas, definidas como aquellas proteínas para las cuales además de contar con una estructura resuelta a menos de 3.5 Å, se cuenta con información acerca de su función específica.

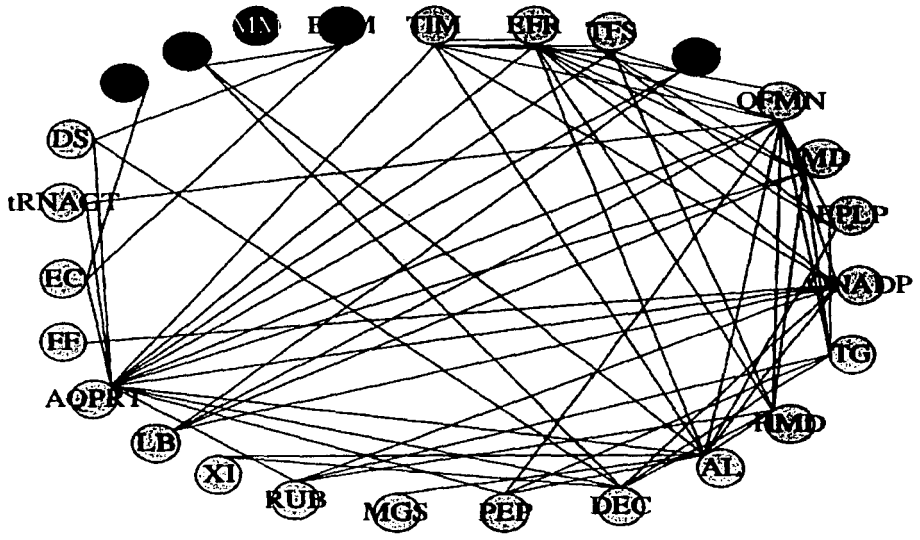


Figura 2. Se ilustran las homologías detectadas entre las 25 de las 26 súper familias de barril TIM. Se presentan en fondo negro las súper familias para las cuales no se había detectado homología previamente y en fondo verde la súper familia para la cual no se encuentra homología alguna Las abreviaciones empleadas son las siguientes: Triosa fosfato isomersa-TIM; Enzimas que unen ribulosa fosfato-ERF; Tiamina fosfato sintasa-TFS; Piridoxin 5' fosfato-P5F; Óxidorreductasa que unen FMN-OFMN; Inosin monofosfato deshidrogenasa-IMD; Enzimas que unen PLP-EPLP; Óxidoreductasas que unen NADP-ONADP; Transglicosidasas-TG; Hidrolasas metal-dependiente-HMD; Aldolasa-AL; Dominio enolasa C-terminal-DEC; Fosfoenolpiruvato-PEP; Malato G sintasa-MGS; Rubisco-RUB; Xilosa isomersa-XI; Luciferaza semejante a la de bacterias-LB; Ácido quinolínico PR-transferasa-AQPRT; Fosfolipasa fosfatidilinositol –FF; Enzimas dependientes de cobamlamina vitamina B12-EC; tRNA-guanín transglicosidasa-tRNAGT; Dihidropteroato sintasa-DS; Uroporfirinogeno dsecarboxilasa-UD; Metilenetetrahidrofolato reductasa-MR; 5. Monometilamin metiltransferase-MM; Betain-homocisteina S-metiltransferasa-BHM.

Existe un grupo de diez súper familias de proteínas en torno a las cuales se agrupan claramente el resto, que son: las TIM, las óxidorreductasas que unen FMN, las inosin monofosfato deshidrogenasa, las óxidorreductasas que unen NADP, las transglicosidasas, las enzimas que unen PLP, las hidrolasas metal-dependiente, las aldolasas, las dominio enolasa C-terminal y las ácido quinolínico PR-transferasa.

La mayoría de las 26 súper familias comprenden solo barriles α/β_8, sin embargo, hay tres familias que poseen siete elementos beta, la súper familia de las Transglicosidasas una endonucleasa la súper familia ácido quinolínico PR-transferasa y la Luciferina semejante a la de bacterias. La súper familia Monometilamin transferasa posee, además de la estructura básica del barril dos pares de hojas betas accesorias.

En las 26 súper familias hay 61 números diferentes de E.C. los cuales comprenden las clases de la uno a la cinco, como se muestra en la Figura 3, de éstos, el 85% está involucrado en el metabolismo energético, metabolismo de macromoléculas o en el de pequeñas moléculas. Las únicas excepciones las constituyen una proteína relacionada con la vía informativa del DNA endonucleasa IV, 1qtw y otra proteína relacionada con la ruta informativa del RNA tRNA-guanín transglicosidasa, 1wkf.

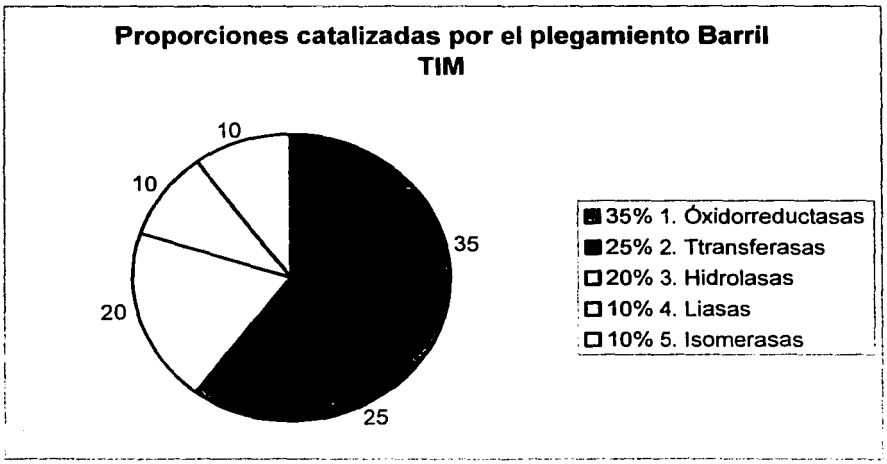


Figura 3. Se muestran las proporciones de las actividades catalíticas llevadas a cabo por los barriles TIM.

Doce familias enlazan a un cofactor tal como FMN, NADP, PLP o metales divalentes. Todos los cofactores y sustratos son enlazados al carbono terminal de las hojas β en todas las familias.

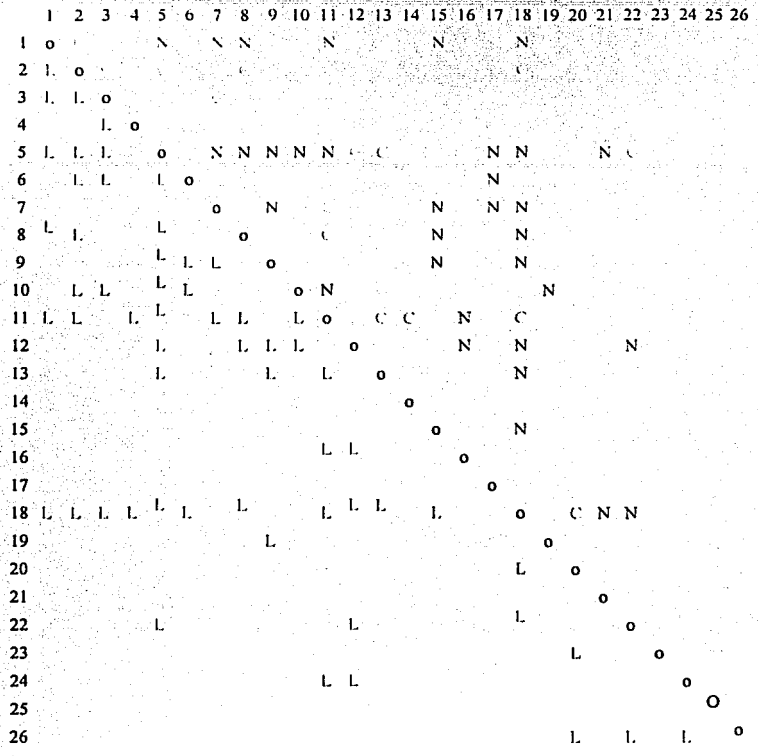
Tabla 3. Enzimas para las cuales no se había detectado homología alguna con el resto de los barriles TIM.

Proteína	ID Å	súper familia	Función	Referencia
Piridoxin 5' fosfato Sintasa de *Escherichia coli*	1hol 2.20*	Piridoxin 5' fosfato	ATP + piridoxal = ADP + piridoxal 5'-fosfato. Cofactor: zinc o magnesio	Garrido-Franco [29]
Uroporfirinogeno dsecarboxilasa del Humano	1uro 1.80	Uroporfirinogeno dsecarboxilasa	Uroporfirinogeno-III = coproporfirinogeno + 4 CO2. Ruta: biosíntesis de Porfirin	Whitby [30]
Metilenetetrahidrofolato reductasa de *E. coli*	1b5t 2.50	Metilenetetrahidrofolato reductasa	5-metiltetrahidrofolato + aceptor = 5,10-metilenetetrahidrofolato + aceptor reducido Cofactor: FAD. Ruta: Biosíntesis de metionina. Es un homo tetrámero	Guenther[31]
Betain-homocisteína S-metiltransferasa	1lt8 2.05	Betain-homocisteina S-metiltransferasa	Trimetilamonioacetato + L-homocisteína = dimetilglicina + L-metionina. Esta reacción también es requerida para la oxidación irreversible de la colina Cofactor: Zinc Ruta: Regulación del metabolismo de homocisteína	Evans[32]

Å Resolución en Amstrongs

En el presente trabajo se muestra que es posible inferir homología entre cuatro súper familias para las cuales no se había reportado antes homología alguna, éstas son: la piridoxin 5'fosfato, la uroporfirinogeno descarboxilasa, la metilenetetrahidrofolato reductasa y la betain-homocisteina S-metiltransferasa. Cabe aclarar que las últimas dos súper familias mencionadas, no se habían reportado hasta diciembre del año pasado. Es importante resaltar que la súper familia metilenetetrahidrofolato reductasa se encuentra involucrada en la síntesis de metionina a partir de homocisteína, y que se detecta homóloga de la familia betain-homocisteina S-metiltransferasa, previamente se habían reportado como homólogas las súper familias de las síntesis de triptofano y de histidina, con lo cual, se ha establecido homología entre cuatro súper familias involucradas en la síntesis de aminoácidos.

Tabla 4. Homologías detectadas entre las 26 súper familias de barril TIM. Los números de las súper familias se encuentran en el mismo orden establecido en la Tabla 1. Aquellas homologías encontradas en los trabajos de Copley *et al.* se muestran con una C, mientras que las detectadas por Nagano *et al.* con una N y las del presente trabajo con una L.



En el trabajo de Copley *et al.* se logró establecer homología entre 12 súper familias de barril TIM, mientras que Nagano *et al.* lo hacen para 21 de ellas. Ellos logran obtener un grupo sólido entre las súper familias de las aldolasas clase I y II, dihidropteroato sintasa, enzimas que unen PLP, las fosfatidilinositol fosfolipasa C, las enzimas que unen fosfato, fosfoenolpiruvato, la triosa fosfato isomersa, oxidorreductasa que unen FMN, ácido quinolínico PR-transferasa, Rubisco y la súper familia de las enolasas. Sin embargo, entre las súper familias transglicosidasas, hidrolasas metal-dependiente, xilosa isomersa y tRNA-guanín transglicosidasa se encuentran compartidos los residuos catalíticos y los residuos de unión a fosfato, lo cual podría apoyar que se unieran de manera distante al grupo

mencionado. En el presente trabajo, para estas familias se detecta homología mediante el análisis realizado con PSI-BLAST.

Aún cuando las súper familias de barril TIM parecen haber divergido lo suficiente en la secuencia, como para no detectar homologías entre ellos a partir de alineamientos simples, con métodos sensibles para detectar homologías que no son evidentes en la secuencia, es posible encontrar homología entre 25 de ellas. Esto nos podría sugerir que los barriles TIM provienen de un ancestro común, sin embargo, la penúltima súper familia, la monometilamin metiltransferasa, no resulta ser homóloga de ninguna de ellas. La estructura de dicha enzima posee elementos de hojas beta repetidas, lo cual podría influir en la evolución de la secuencia, resultando en una rápida divergencia, pues al realizar el PSI-Blast, sólo logra reconocer al plegamiento Rossman. Esta enzima cataliza varias reacciones, resaltadas en la Lista 1, así como varios pasos dentro de una misma ruta, como en el caso de la biosíntesis de la ubiquinona, en donde cataliza tres pasos continuos.

De forma tal que, aún cuando las 26 súper familias de barriles TIM, hasta ahora reconocidas, parecen haber divergido en sus secuencias de aminoácidos, evitando que sea posible detectar homología entre ellas, a partir de alineamientos de secuencia con métodos poco sensibles, como BLAST o FASTA; el uso de métodos más sensibles, como PSI-BLAST, a la par del criterio de homología transitiva, nos ha permitido detectar relaciones monofiléticas entre 25 de las 26 súper familias.

Por otro lado, el hecho de que las proteínas de la súper familia de las monometilamin metiltransferasas, no han podido ser ligadas por métodos de alineamiento, incluyendo PSI-BLAST, al resto de las súper familias con barriles TIM (Farber, Copley, Nagano y este trabajo), pudiera explicarse sobre la base de dos posibles causas. La primera es que 25 súper familias tengan un origen común y las monometilamin metiltransferasas uno independiente, lo cual es difícil de evidenciar o descartar a ciencia cierta, no solo en este caso, sino en cualquier tipo de proteínas, con plegamientos similares, en los que no se encuentren evidencias de homología a nivel de secuencia.
La segunda es que las secuencias de aminoácidos de las monometilamin metiltransferasas han divergido tanto, del resto de las súper familias, que incluso por los métodos aquí empleados, no se puedan detectar indicios de homología, si los hubiere. Esto es probable en

el sentido de que la estructura de las monometilamin metiltransferasas poseen elementos de hojas beta repetidas, a manera de inserciones que interrumpen el plegamiento prototípico de barril TIM, lo cual podría influir en la evolución de la secuencia, resultando en una rápida divergencia, y de ello que al realizar el PSI-BLAST, no logremos ligarla con otros tipos de barriles TIM, de hecho, las proteínas que mas se parecen en a las monometilamin metiltransferasas, según los resultados de PSI-BLAST, son proteínas con plegamientos tipo Rossman, aún cuando los elementos "insertados" en el dominio barril-TIM las monometilamin metiltransferasas han sido eliminados para su comparación por los métodos aquí empleados.

Las monometilamin metiltransferasas catalizan varias reacciones, resaltadas en la Lista 1, así como varios pasos dentro de una misma ruta, como en el caso de la biosíntesis de la ubiquinona, en donde cataliza tres pasos continuos. Lo que nos permite sugerir que si los loops de las monometilamin metiltransferasas actúan como elementos importantes en la selección de los diferentes tipos de sustratos y ello han sido un factor importante para una divergencia a nivel de secuencia muy alta, nos podemos inclinar por la segunda opción descrita arriba; es decir, que las monometilamin metiltransferasas pudieran ser homólogas al resto de los barriles TIM, pero que hasta la fecha, por los métodos aquí empleados no podamos detectar señales de ello. Por lo cual es necesario extender el espacio de búsqueda, de manera tal que se pudieran detectar proteínas homólogas transitivas entre las monometilamin metiltransferasas y el resto de los barriles TIM.

Lista 1. Reacciones catalizadas por la enzima Monometilamin metiltransferasa

Metabolismo

- Vitaminas y cofactores
- Lípidos andrógenos y estrógenos
- Aminoácidos histidina, triptofano, tirosina
- Otros aminoácidos aminofosfonato y selenoaminiácido

Biosíntesis

Ubiquinona

Biodegradación de xenobióticos degradación de nitrobenceno

Todas las proteínas analizadas son enzimas o están claramente relacionadas a una enzima y se encuentran involucradas en el metabolismo molecular o energético, el cual está considerado como la más antigua de las funciones biológicas. Con el análisis realizado se puede apoyar la hipótesis de Farber *et al.* acerca de la evolución divergente de este plegamiento y el caso de la súper familia 25, monometilamin metiltransferasa, la podríamos sugerir como la enzima cuya evolución ha sido muy temprana y cuyas presiones de selección han sido diferentes a las del resto del plegamiento, dados los elementos estructurales las hojas β adicionales que posee y el hecho de que sea una enzima con varios sustratos.

GLOSARIO

ASTRAL: El compendio ASTRAL, proporciona varias bases de datos y herramientas que ayudan en el análisis de las estructuras de las proteínas, particularmente a través del uso de sus secuencias. El "score" SPACI, incluido en el sistema, resume las características generales de las estructuras. Su dirección es la siguiente: http://astral.stanford.edu/

BASES DE DATOS: Comprende toda la información disponible, ya sea de nucleótidos o de proteínas, a una fecha determinada, reunida en un mismo sitio. Un concepto dominante en comparar bases de datos es la aplicación del de redundancia. Muchas bases de datos intentan ser "no-redundantes". Desafortunadamente, los datos biológicos son demasiado complejos para que tenga lugar una definición simple de la redundancia. ¿Se puede decir que dos alelos son redundantes? ¿Dos isozimas en el mismo organismo? Por lo tanto, cada base de datos "no-redundante" tiene su propia definición de la redundancia. Algunas hacen cierto uso de medidas automatizadas, mientras que otras utilizan el desecho manual. Otras bases de datos no procuran ser no-redundantes, sino que sacrifican algo esta meta en favor de asegurar lo completo.

BLAST: Basic Local Alignment Search Tool El algoritmo alinea regiones cortas de la secuencia desconocida con regiones encontradas en la base de datos. La fase inical de la búsqueda consiste en identificar fragmentos parecidos en la base de datos. Luego se alinean los aminoácidos de los fragmentos con la secuencia desconocida, si un aminoácido de la secuencia desconocida se encuentra ubicado exactamente en la misma posición de la

secuencia de la base de datos, se le adjudica un punto positivo. En el caso de que la similitud e identidad fuesen buenas pero no perfecta, entonces se le adjudica un puntaje más bajo. Cuando no existe una correspondencia buena y perfecta entre los aminoácidos se le adjudica un puntaje negativo. La suma de los puntajes es utilizada para determinar el grado de similitud.

Las secuencias con puntajes altos son denominadas "*high-scoring segments pairs, HSPs*". El programa trata de extender el mejor HSP aquellos con los mayores puntajes, los más parecidos alineando en ambas direcciones hasta que la secuencia se acabe o el mismo deje de ser estadísticamente significativo. Durante los procesos de búsqueda y extensión se hace uso de matrices de substitución. Entonces las secuencias reportadas serán aquellas que posean los puntajes totales más altos "*maximal-scoring segment pair, MSP*".

La longitud del segmento inicial a ser identificado se especifica por el valor W "*Wordlength*". El algoritmo BLAST solamente intenta extender el fragmento alineado si existe una correspondencia perfecta por los W aminoácidos continuos. El valor por omisión del programa Blast es de 11 letras, buscará en la base de datos hasta que consiga fragmentos de 11 letras de largo que sean exactos a la secuencia desconocida para luego proceder a la extensión hacia ambos lados. Una longitud de 11 letras es suficiente para excluir homólogos que divergen moderadamente y por consiguiente también excluirá aquellos que puedan haber sido escogidos por azar.

CATH: Agrupa a las proteínas de acuerdo al orden jerárquico de sus iniciales en cuatro niveles importantes, ClaseC, ArquitecturaA, TopologíaT y súper familia homólogo h El nivel de clase, derivado del contenido de la estructura secundaria, se asigna para más del 90% de estructuras de la proteína automáticamente. La arquitectura, que describe la orientación gruesa de estructuras secundarias, independientemente de conectividades, se asigna, actualmente, de manera manual. El nivel de la topología agrupa las estructuras según sus conexiones y el número de estructuras secundarias. Las súper familias homólogas agrupan a las proteínas con base a las estructuras altamente similares y a las funciones. Las asignaciones de estructuras a las familias de topología y a las súper familias homólogas son hechas por comparaciones de la estructura primaria, secundaria y terciaria.

CORA: Son una serie de programas para alineamiento múltiple y análisis estructural de familias para identificar las posiciones consenso y capturar sus características estructurales

más conservadas por ejemplo, accesibilidad de los residuos, ángulos de torsión, geometría global descrita mediante los contactos entre residuos.

GEN: Unidad física fundamental de la herencia cuya existencia se puede confirmar por variantes alélicas y que ocupa un locus cromosómico concreto. Secuencia de DNA que codifica para un polipéptido o un producto determinado, como puede ser un marco de lectura abierto ORF, por sus siglas en inglés rRNA o un tRNA.

GEN BANK: Es una colección de todas las secuencias disponibles del DNA *Nucleic Acids Research* 2002 Jan 1;301:17-20. Hay aproximadamente 22.617.000.000 bases de DNA en 18.197.000 registros de secuencia con fecha de agosto del 2002. GenBank es parte de "Nucleotide Sequence Database Collaboration", que engloba al Banco de Datos de DNA de Japón DDBJ, al Laboratorio Europeo de Biología Molecular EMBL, y al GenBank del NCBI. Estas tres organizaciones intercambian datos diariamente.

GENOMA: Conjunto de información genética de un organismo determinado. Contiene, por tanto, la información necesaria para que las células sinteticen todos los componentes del organismo.

HOMOLOGÍA TRANSITIVA: Siguiendo la misma lógica aplicada en la propiedad de transitividad de los números reales

En donde, si A=B y B=C.: A=C podemos encontrar homologías transitivas entre proteínas que de otra manera no hubiese resultado evidente.

IMPALA: Es un programa diseñado para complementar el procedimiento de comparación de una secuencia problema con una base de datos de matrices de registro posición–dependiente PSSMs generadas por PSI-BLAST.

NCBI: Establecido en 1988 como un recurso nacional para la información molecular de la biología, el Centro Nacional de Biotecnología NCBI, por sus siglas en inglés crea bases de datos públicas, conduce la investigación en biología de cómputo, desarrolla las herramientas del software para analizar datos del genoma y disemina la información biomédica.

PDB: El Banco de Datos de Proteínas PDB, por sus siglas en inglés es la colección sobre algunos datos disponibles con respecto a estructuras de proteínas, resueltas con difracción de rayos X, datos de resonancia nuclear magnética, ácidos nucleicos y otras macromoléculas biológicas. Sus entradas pueden clasificarse en

- Estructuras de proteínas, las cuales pueden incluir cofactores, sustratos, inhibidores u otros ligandos, incluyendo ácidos nucleicos.
- Estructuras de oligonucleótidos o ácidos nulceicos.
- Modelos hipotéticos de estructuras de proteínas.

PLEGAMIENTO DE PROTEÍNAS: Proceso mediante el cual la proteína sintetizada en la célula adquiere una conformación definida en el espacio que le permite reconocer otros componentes celulares y así desarrollar su función característica. La naturaleza de la secuencia de aminoácidos determina las interacciones con el medio citoplasmático, las cuales pueden ser hidrofóbicas o hidrofílicas, tomando en conjunto una conformación característica.

PROTEÍNA: Macromolécula orgánica constituida por aminoácidos unidos por enlaces peptídicos que tiene un papel funcional y además estructural en los procesos vitales de la célula.

PSI-BLAST: Es una versión modificada de BLAST, pero iterativo. Del universo de proteínas que se obtienen con la primer salida, construye un perfil, empleándolo como semilla para realizar una segunda búsqueda, de tal suerte que en esta segunda búsqueda ya no solo se realiza con base en la secuencia original, sino con base en el perfil generado, por lo cual resulta un método sensible a homologías que no son evidentes en la secuencia y que con BLAST no se detectarían.

SCOP: Casi todas las proteínas tienen semejanzas estructurales con otras proteínas y, en algunos de estos casos, comparten un origen evolutivo común. La base de datos de SCOP, creada mediante una inspección manual y sistematizada, empleando una batería de métodos automatizados, proporcionando una descripción detallada y comprensiva de las relaciones estructurales y evolutivas entre todas las proteínas de estructura conocida. Como tal, provee un amplio examen de todos los plegamientos conocidos de la proteína, así como información detallada sobre los parientes cercanos de cualquier proteína particular, y un marco para la investigación y la clasificación futuras. Organiza las estructuras de las proteínas en una jerarquía de acuerdo con su origen evolutivo y a su similitud estructural. Todo ello con base en los dominios de las proteínas más que con toda la proteína completa. Por lo cual, en el nivel más bajo de la jerarquía del SCOP, hay dominios individuales, extraídas a partir de las entradas del PDB. Los grupos de dominios están en conjunto dentro

de familias homólogas. Estos dominios abarcan similitudes en la estructura, función y en la secuencia, lo cual, implica un origen común.

VALOR DE EXPECTANCIA: En una búsqueda de similitud en una base de datos, la probabilidad de que un registro score de alineamiento tan bueno como el encontrado entre una secuencia problema y una secuencia en la base de datos pueda ser encontrado en tantas comparaciones entre secuencias al azar como fue encontrada la secuencia apareada.

Referencias

[1] Chothia, C. 1992. One thousand families for the molecular biologist. Nature. **357.**

[2] Lo Conte Bart Ailey, Tim J. P. Hubbard, Steven E. Brenner, Alexey G. Murzin, and Cyrus Chothia, C. 2000. SCOP: a structural classification of proteins database. Nucl. Acids. Res. **28** 257-259.

[3] Lo Conte, L., Hubbard, T., Brenner, S., Murzin, A., Chothia, C. 2000. SCOP: a structural classification of proteins database. Nucl. Acids. Res. **28** 257-259.

[4] Orengo, A., Jones, S., Jones, D., Swindells, M., Thornton, J. 1997. CATH: a hierarchic classification of protein domain structures. Structure **5** 1093-1108.

[5] Dayhoff. 1976. The origin and evolution of Protein superfamilies. Fed. Proc. **35** 2132-2138.

[6] Koonin, E., Wolf, Y., Karev, G. 2002. The structure of the protein universe and genome evolution. Nature. **240** 218-223.

[7] Murzin, A., Lo Conte, L., Andreeva, A., Howorth, D., Ailey, B., Brenner, S., Hubbard, T., Chothia, C.1995. SCOP: a structural classification proteins database for the investigation of sequences and structure. J. Mol. Biol. **247** 536-540

[8] Riley, M., Labedan, B. 1997. Protein evolution viewed through *Escherichia coli* protein sequences: Introducicing the notion of a structural segment of homology, the module. J. Mol. Biol. **268** 857-868.

[9] Martin, A., Orengo, C. Hutchinson, E. Jones, S. Karmirantzou, M. Laskowki, R. Mitchell, J. Taroni, C. Thornton, J. 1998. Protein folds and functions. Structure. **6** 875-884.

[10] Babbitt, P., Gertl, J. 1997. Understanding enzyme superfamilies. J. Biol. Chem. **272** 30591-30594

[11] Webb, E. 1992. Enzyme Nomenclature 1992. Recommendations of the Nomenclature Committee of the international Union of Biochemistry and Molecular Biology, Academic Press, New York.

[12] Branden, C., Tose John. 1991. Introduction to protein structure. Garland Publishing, Inc. New York and London. pp. 43-44

[13] Farber, G., Petsko, G. 1990. The evolution of $(\beta\alpha)_8$ barrel enzymes. Trends Biochem. Sci. **15** 228-234

[14] Kimura, M. 1968. Evolutionary rate at the molecular level. Nature. **217** 624-626.

[15] Kimura, M. 1979. The neutral theory molecular evolution. Scientific American. **241** 98-126.

[16] Needleman, S., Wunsch, C. 1970. A general method applicable to the search for similarities in the amino acid sequences of two proteins. J. Mol. Biol. **48** 444-453.
[17] Smith, T., Waterman, M. 1981. Identification of common molecular subsequences J. Mol. Biol. **147** 195-197.
[18] Lipman, D., Pearson ,W. 1985: Rapid and sensitive protein similarity search, Science, **227** 1435-1444.
[19] Karlin, S., Altschul, S. 1990. Method for assessing the Statical Significance of Molecular Sequence Features by using General Scoring Schemes Proc. Natl. Acad. Sci. USA **87**:2264-2268.
[20] Park., J. Teichmann, S., Hubabard, T., Chothia, C. 1997. Intermediate sequences increase the detection of homology between sequences. J. Mol. Biol. **273** 349-354.
[21] Copley, R., Bork, P. 2000 Homology among $(\beta\alpha)_8$ barrel: implications for the evolution of metabolic pathways. J. Mol. Biol. **303** 627-640.
[22] Brenner, S. Chandonia, J., Lo Conte, L., Walker, N., Koehl, P., Levitt, M. 2000. The ASTRAL compendium for protein structure and sequence analysis. Nucl. Acids Res. **28** 254-256.
[23] Altschul, S., Wolf, Y., Ponting,C., Koonin, E., Aravind, L., Schäffer, A.1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids. Res. **25** 3389-3402.
[24] Russel, T., Barton, D. 1992. Multiple Protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. Proteins: Struct. Funct. Genet. **14** 309-323.
[25] Nagano, N., Orengo, C., Thornton, J. 2002. One fold with Many Functions: The Evolutionary Relationships between Tim Barrel Families Based on their Sequences, Structures and Functions. J. Mol. Biol. **321** 741-765.
[26] Schaffer, A., Altschul, S., Wolf, Y., Ponting,C *et al.* 1999. IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics, **15** 1000–1011.
[27] Orengo, C. 1999. CORA-topological fingerprints for Protein structurals families. Protein Sci. **8** 699-715.
[28] Lang, D., Thoma, R,. Henn-Sax, M., Sterner, R., Wilmanns., M. 2000. Structural evidence for evolution of the β/α barrel scaffold by gene duplication and fusion. Science **289** 1546-1550.
[29] Garrido-Franco, M., Laber, B., Huber, R., Clausen, T. 2002. Enzyme-ligand complexes of pyridoxine .5'phosphate synthase: implications for substrate binding and catalysis. J. Mol. Biol. **321** 601-612.
[30] Whtiby, M 1998. Crystal structure of human uroporphyrinogen decarboxylase. EMBO. J. **17** 2463-2471.
[31] Guenther, B., Sheppard, C., Tran, P., Rozen, R., Matthews, R., Ludwig, M., 1999. The structure and properties of methylenetetrahydrofolate reductase from Escherichia coli suggest how folate ameliorates human hyperhomocysteinemia. Nat Struct Biol. **6** 359-65.
[32] Evans, J., Huddler, D., Jiracek, J., Castro, C., Millian, N., Garrow, T., Ludwig, M. 2002. Betaine-homocysteine methyltransferase: zinc in a distorted barrel. Structure (Camb). **10** 1159-71.

Apéndice

Script I

```
FILE="$1"
TEMPFILE="temporal.awk"

grep "|" $FILE > $TEMPFILE

awk ' BEGIN {digits = "^[0-9]+$"; alfa= "^[0-9e'-']+$" }
! /^>/  && $(NF-1) ~ digits && $NF ~ alfa' $TEMPFILE
```

Script II

```
for i in $(ls $1)
do
for j in $(ls $1)
do
cat $1/$i | awk -F '|' '{print $2}' | sed "s/ //g" >> /tmp/temp1
for x in $(cat /tmp/temp1)
do
grep $x $1/$j >> $2/$i-$j
done
rm /tmp/temp1
done
done
```