UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

INSTITUTO DE BIOTECNOLOGÍA

ANÁLISIS DE PERFILES DE ENTROPÍA EN FAMILIAS ESTRUCTURALES DE PROTEÍNAS

Tesis que para obtener el título de

Doctor en Ciencias Bioquímicas

Presenta

Fidel Alejandro Sánchez Flores

CUERNAVACA, MORELOS **MAYO 2007**

RECONOCIMIENTOS

DIRECTOR DE TESIS

Dr. Lorenzo Patrick Segovia Forcella	IBT/UNAM

MIEMBROS DEL COMITÉ TUTORAL

Dra. Nina Pastor Colón	Fac. Ciencias/UAEM
Dr. Alejandro Garciarrubio Granados	IBT/UNAM
Dr. Gabriel Moreno Hagelsieb	Universidad de Waterloo/Canada

MIEMBROS DEL COMITÉ TUTORAL AMPLIADO

Dra. Nina Pastor Colón	Fac. Ciencias/UAEM
Dr. Alejandro Garciarrubio Granados	IBT/UNAM
Dr. Enrique Merino Perez	IBT/UNAM
Dr. Enrique Rudiño Piñera	IBT/UNAM
Dr. Bruno Contreras Moreira	CCG/UNAM

MIEMBROS DEL JURADO

Dr. Enrique Merino Perez	IBT/UNAM
Dr. Enrique Rudiño Piñera	IBT/UNAM
Dr. Bruno Contreras Moreira	CCG/UNAM
Dr. Alejandro Sosa Peinado	Fac. Medicina/UNAM
Dr. Lorenzo Patrick Segovia Forcella	IBT/UNAM

BECAS OTORGADAS POR:

CONACYT
DGEP-UNAM

APOYO TECNOLÓGICO:

Este trabajo se realizó con el equipo de cómputo Cluster Sputnik II donado por El Macroproyecto de Tecnologías de la Información y la Computación de la Universidad Nacional Autónoma de México.

AGRADECIMIENTOS

A TODA mi familia. Lado paterno, lado materno, hoy ya no hay divisiones para mi. Ausentes y presentes todos son y serán siempre parte de quien soy.

A mis mentores: Julieta Rubio, Clementina Castro, Brenda Valderrama y Lorenzo Segovia. Cada uno aportó con gran esfuerzo y dedicación a mi formación y hoy después de 10 años les digo GRACIAS.

A la Universidad Nacional Autónoma de México por ser mí segunda casa.

Al Instituto de Biotecnología. No he encontrado mejor ambiente de trabajo y sobre todo su gente es lo más valioso.

A la gente del laboratorio. Aquí si los voy a mencionar uno a uno: Arcadio, Viviana, las Marianas, Iliana, Adriana, (puras ianas…) Areli, Ernesto, Dago, Lorena, Yinna, Haven, Javier. Gracias por aguantarme durante todo este tiempo, sobre todo los que llevan más años haciéndolo.

A todos mis amigos, del DF y Cuernavaca. Espero seguir contando con ustedes siempre.

A Jimena, por siempre haberme apoyado durante el tiempo que se pudo. Tú me ayudaste a cambiar.

A Anilu, siempre estuviste ahí y nunca has dejado de creer en mí. GRACIAS.

Y a todas aquellas personas que me conocen, GRACIAS. Aunque no aparezca su nombre en esta página, son participes de éste mi gran logro

GRACIAS, GRACIAS, GRACIAS.

INTRODUCCIÓN

Generalidades

Las **proteínas** moléculas muy importantes para los seres vivos, ya que realizar la mayor parte de las actividades biológicas. Son polímeros compuestos por 20 diferentes **aminoácidos,** aunque no necesariamente los encontramos a todos formando parte en ellas. Al orden o la secuencia de los aminoácidos que forman la cadena peptídica se le conoce como la **estructura primaria** de la proteína. Este es el primer nivel de organización en las proteínas y cualquier nivel de organización subsecuente, está determinado en gran parte por la estructura primaria [1].

El siguiente nivel de organización que se observa son estructuras locales a lo largo de la cadena peptídica y que se forman por las interacciones no covalentes entre los aminoácidos vecinos en ciertas regiones de la proteína. A estas conformaciones locales se les denominan **estructuras secundarias** y existen 3 arreglos básicos que son: **hélice alfa, láminas** u **hojas beta** y las **"vueltas"** o **giros**. Cualquier otro tipo de conformación es una variante de alguna de estas estructuras. A su vez, estas estructuras secundarias pueden formar patrones que se repiten dentro de la estructura e incluso se observan entre proteínas que no están relacionadas. A este tipo de arreglos se les denominan **estructuras supersecundarias** [1, 2].

Las estructuras secundarias que se observan en las proteínas pueden ser mayoritariamente de un tipo o bien, una mezcla de ellos. Estas estructuras hacen contacto entre si mismas formando un nuevo nivel de organización, donde en la mayoría de los casos, se forma un glóbulo y el trazo particular en el espacio que describe la cadena de aminoácidos, conforma el tercer nivel de organización llamado **estructura terciaria** [1].

Finalmente, podemos observar un nivel más de organización llamado **estructura cuaternaria,** que se observa cuando una proteína está conformada por más de una cadena polipeptídica y estas cadenas interactúan. A cada una de las cadenas se le denomina como una **subunidad** de la proteína y éstas pueden ser idénticas o no idénticas. Las interacciones que se pueden observar entre las distintas cadenas son prácticamente las mismas que se observan en la estructura terciaria. La estructura cuaternaria regula muchas funciones en las proteínas, así como también afecta la estequiometría de las reacciones que se llevan a cabo. En algunos casos, las subunidades pueden llevar a cabo la función sin necesidad de las otras cadenas, sin embargo existen casos donde los aminoácidos que interaccionan con los sustratos, se encuentran en cadenas distintas y el ensamblado de las distintas cadenas es necesario para llevar a cabo el reconocimiento de sustrato y la función. En la Figura 1, se muestra esquemáticamente los diferentes niveles de organización mencionados[1, 2].

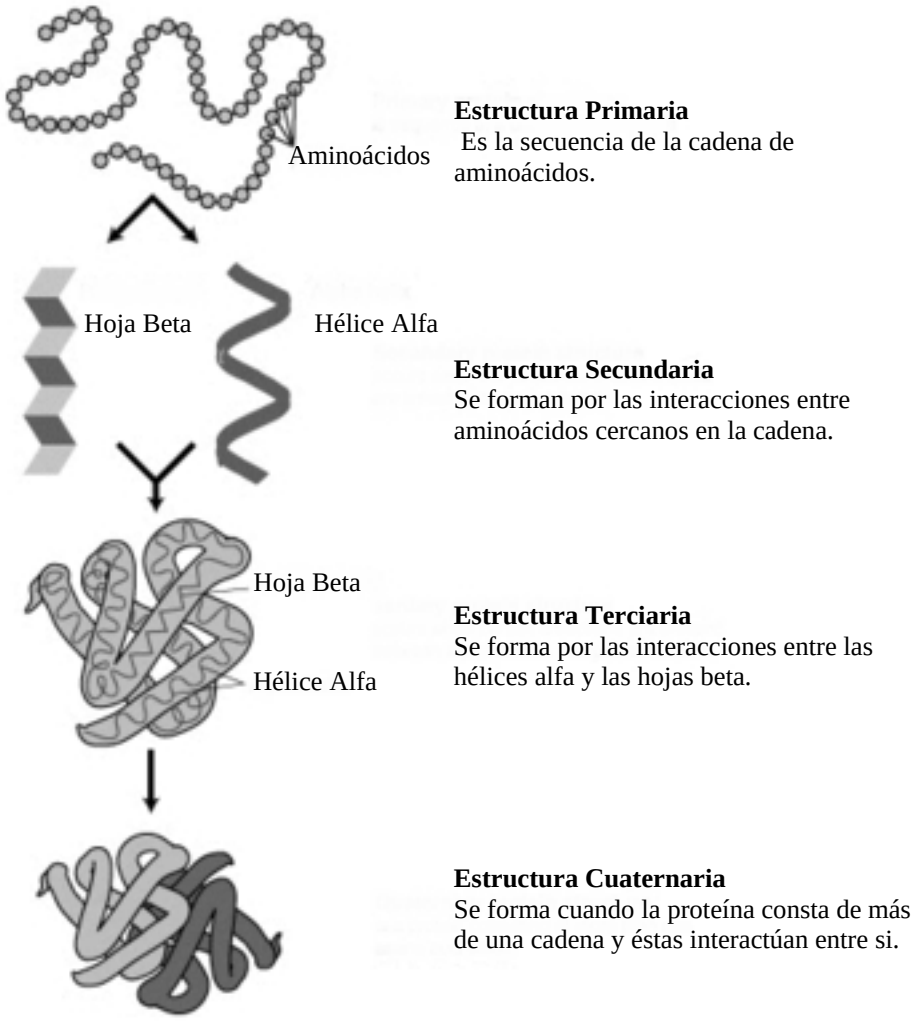


Figura 1.- Representación de los diferentes niveles de estructuración en las proteínas.

El plegamiento

La estructura terciaria o **plegamiento**, es un tema de estudio importante para entender la función y evolución de las proteínas. A *grosso modo*, la forma final que adoptará una proteína será casi esférica o globular. Existen otras formas que pueden adoptar las proteínas como las estructurales (proteínas fibrilares), proteínas de membrana o proteínas de secreción, sin embargo la forma globular es la mas estudiada.

Cuando una proteína se sintetiza en la célula, se encuentra en un estado desplegado del cual tiene que pasar a una conformación estable y que termodinámicamente es un estado de menor energía. En un modelo clásico, observamos que una proteína pasa de un estado no plegado con cierto nivel de **energía libre de Gibbs**, a un estado de mayor energía el cual se conoce como **estado de transición** para llegar finalmente a un estado plegado que es de menor energía. Gracias a estos estudios de termodinámica y cinética de plegamiento, podemos saber el camino que siguen algunas proteínas para llegar a una forma estructurada y la contribución relativa de cada residuo a la estabilidad para adquirir el estado plegado[3].

Actualmente, sabemos que las proteínas no pasan por todas las conformaciones posibles, ya que tomaría una cantidad de tiempo casi infinita tan solo para una proteína de 100 residuos. Todo parece indicar que el mecanismo por el cual una proteína se pliega implica una secuencia de varios pasos. Posiblemente el modelo que mejor explica el mecanismo de plegamiento de las proteínas es el de **nucleación-condensación**[4] que es una combinación de dos mecanismos propuestos con anterioridad: El modelo de marco (**framework**) donde primero se forman las estructuras secundarias, las cuales interaccionan favoreciendo la formación de la estructura terciaria[5]. El segundo modelo es el de **colapso hidrofóbico**, donde como el nombre lo indica, la proteína se compacta en un espacio globular definido donde el número de interacciones posibles se reduce con lo cual la proteína alcanza su topología final [6].

En una cadena polipeptídica podemos observar regiones que pueden plegarse de manera independiente y pueden presentar o no el mismo tipo de plegamiento. A estas unidades se les conocen como **dominios estructurales**. Cabe resaltar que este **mosaico** de plegamientos que se observan en una misma cadena, contienen parte de la historia evolutiva de la proteína que obedecen a ciertas reglas como duplicaciones internas a nivel del gene que las codifica o bien, fusión de diferentes genes codificantes [2].

La **topología** de la proteína es la descripción detallada de los arreglos especiales determinados por la estructura primaria, de como se dan los contactos necesarios para conformar una molécula bien empacada, donde las cadenas laterales de los aminoácidos que forman las estructuras secundarias con propiedades hidrofóbicas se colapsan hacia adentro; las partes hidrofílicas quedan expuestas al medio acuoso; se forman las cavidades donde el sustrato se une y los aminoácidos involucrados en la catálisis quedan orientados de manera en que puedan llevar a cabo una reacción determinada.

Usualmente, se reconocen 4 topologías básicas en las proteínas, dependiendo del tipo de estructuras secundarias que las conforman [2]:

- **Mayoritariamente alfa**: Los residuos de la cadena polipeptídica se encuentran formando estructuras tipo alfa hélice primordialmente. Estas hélices se encuentran en contacto unas con otras. El patrón regular básico o estructura supersecundaria de esta topología es el sube-y-baja (**up-and-down**), que son dos hélices consecutivas, adyacentes y antiparalelas, que se pueden repetir varias veces a lo largo de los dominios para formar paquetes de hélices como los observados en los plegamientos tipo **3-helix bundle** o **4-helix bundle** (Fig. 2)**.** Dependiendo la conectividad entre las estructuras secundarias, se pueden observar combinaciones tanto de hélices paralelas como de antiparalelas.

- **Mayoritariamente beta**: Como su nombre lo indica, son proteínas conformadas solamente por hojas beta, ya sean paralelas, antiparalelas o bien una mezcla de ambas. Aquí también podemos observar una patrón básico llamada meandro (**meander**), que son dos hojas beta consecutivas, adyacentes y antiparalelas (Fig. 3 A y B). Debido a la forma que describe en el espacio, se hace analogía con los meandros o curvas que forman cuando un río da la vuelta. La llave griega o **greek-key** es un ejemplo de motivo o estructura supersecundaria que se observa con gran frecuencia dentro de las proteínas mayoritariamente beta. Debido a estos patrones regulares, este tipo de proteínas adoptan ya sea forma de **barril** o bien forma de **sándwich** (Fig. 3 C y D).

- **Alfa / Beta**: Son plegamientos con una mezcla de hélices alfa y hojas beta, donde ambos tipos de estructura secundaria están en continuidad y se encuentran haciendo contacto entre ellas. Dentro de este tipo de topologías encontramos una gran diversidad ya que la conectividad y sentido de cada una de las estructuras secundarias (paralelas o antiparalelas), así como el número de hélices u hojas betas que podemos encontrar a lo largo de un dominio estructural, da como resultado un gran número de combinaciones posibles (Fig. 5).

- **Alfa + Beta**: Son plegamientos que constan tanto de estructuras secundarias tipo alfa hélice y hojas beta, sin embargo las hélices se agrupan en una región de la proteína sin hacer contacto con las hojas beta (Fig. 6).

- **Proteínas intrínsecamente desordenadas**: Este tipo de proteínas forman una clasificación aparte ya que son muy diferentes de las proteínas con un plegamiento definido. Estudios recientes muestran que estas proteínas a pesar de que presentan regiones donde se observan "vueltas azarosas" (**random coils**), desempeñan funciones de gran importancia en la célula, a nivel estructural, señalización, regulación y acoplamiento con otras proteínas. Estructural y evolutivamente, este tipo de proteínas tienen características muy distintas donde su flexibilidad y carencia de una estructuración bien definida, les permite llevar a cabo funciones muy particulares. A diferencia de una proteína desplegada, estas proteínas son estables y se les ha llamado también "plegables" o proteínas "pre-globulares". Presentan una composición sesgada en algunos tipos de aminoácidos como un enriquecimiento en prolina, ácido glutámico, lisina y glutamina; también se observa una muy baja cantidad de aminoácidos aromáticos, cisteína, leucina, isoleucina y asparagina [7].

Figura 2.- Ejemplos de proteínas con topología mayoritariamente alfa. A la izquierda, una proteína con topología 3-helix bundle y a la derecha una con topología 4-helix bundle.

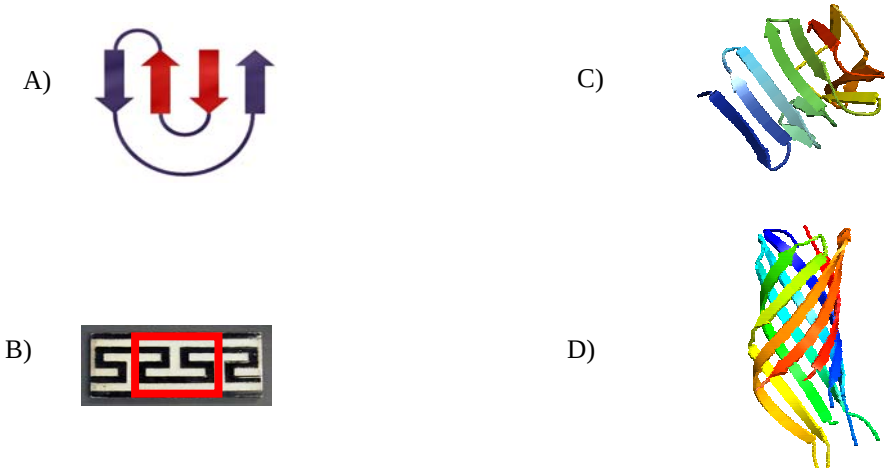


Figura 3.- A) Ejemplo de estructura supersecundaria tipo greek-key. En rojo se observa la estructura supersecundaria básica llamada meandro. B) Motivo griego greek-key al cual hace analogía la estructura supersecundaria del mismo nombre. En rojo se marca la forma de meandro. C) Ejemplo de proteína con plegamiento mayoritariamente beta de tipo sándwich. D) Ejemplo de proteína con plegamiento mayoritariamente beta de tipo barril. En los últimos dos casos, se puede observar la repetición de la forma básica de meandro.

Figura 4.- Ejemplos de proteínas con topologías alfa / beta. A la izquierda una topología tipo a/b/a sándwich y a la derecha un barril $(beta/alfa)_8$.

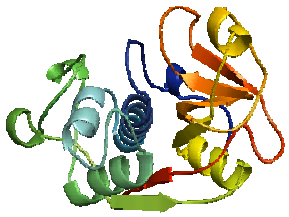


Figura 5.- Ejemplo de una proteína con topología alfa + beta. Como se puede observar, las estructuras secundarias de tipo alfa hélice y las hojas beta se encuentran en regiones distintas en la proteína.

Relación entre secuencia y estructura

Como se había mencionado, la estructura primaria o **secuencia** de la proteína, determina en gran parte los niveles de organización subsecuentes. La predicción de la estructura terciara a partir de la estructura primaria es uno de los grandes retos que se tienen en la biología. Para ello, debemos entender las reglas que determinan el plegamiento de un polipéptido con una secuencia de aminoácidos particular. Dependiendo de los ángulos de torsión que formen entre los aminoácidos, de la distancia a la que se encuentren y de las propiedades de cada aminoácido, se formarán interacciones que dan origen y estabilizan la estructura de la proteína.

Son varias las **interacciones** que se observan en la estructura terciaria de una proteína [1]:

- **Puentes disulfuro:** Son enlaces covalentes entre las cadenas laterales de cisteínas que se encuentran en combinación. Esta interacción es fuerte y solamente se rompe a temperaturas muy altas, pH ácido o por la presencia de un medio o un agente reductor. Este tipo de interacción pone en contacto cercano a las estructuras secundarias, lo cual facilita el empacamiento de la proteína.

- **Interacciones de carga**: Ocurren entre aminoácidos con cadenas laterales con grupos cargados positiva y negativamente (NH_3^+ y COO^-). Estos grupos actúan como partículas cargadas y obedecen la ley de Coulomb, cuya ecuación describe la energía potencial entre dos cargas separadas por una distancia. Sin entrar en detalle del cálculo, encontramos que las variables que afectan estas interacciones son la distancia entre los residuos y la polaridad del medio, ya que los aminoácidos en una región polar, contribuyen en menor cantidad a la estabilidad global de la proteína que cuando se encuentran en una región no polar de la proteína.

- **Puentes de hidrógeno**: Debido a que este tipo de interacciones son las que contribuyen a la formación y estabilidad de las hélices alfa y las hojas beta, es de esperarse que tengan una gran contribución a la estabilidad global de la estructura. Se forman por la interacción de un átomo electronegativo y un átomo de hidrógeno. Los grupos electronegativos como el oxígeno y el nitrógeno los encontramos tanto en la cadena principal o **backbone** de la proteína, como en las cadenas laterales de los residuos y funcionan como grupos aceptores para el puente de hidrógeno. Los átomos de hidrógeno actúan como el grupo donador para el puente de hidrógeno y la distancia a la cual se tienen que encontrar, está en el rango aproximado de 0.26-0.34 nm. Los puentes de hidrógeno se pueden formar entre:

  - átomos de las cadenas laterales de aminoácidos en cadenas distintas
  - átomos en cadenas laterales de aminoácidos y moléculas de agua en la superficie de la proteína
  - átomos en cadenas laterales y los encontrados en el backbone
  - átomos en el backbone y moléculas de agua en la superficie de la proteína
  - átomos en el backbone de dos diferentes aminoácidos.

  Cada caso antes mencionado aporta en diferente magnitud a la estabilidad de la proteína de manera individual, sin embargo es la suma de todas las interacciones lo que contribuye a la estabilidad del plegamiento.

- **Interacciones de van der Waals**: Estas interacciones son de suma importancia para el plegamiento de una proteína se observan entre átomos no cargados y que no están realizando ningún otro tipo de interacción. Surgen por la inducción de momentos dipolares temporales en dichos átomos. Estos momentos dipolares se producen constantemente ya que la densidad de carga de los átomos fluctúa constantemente. El dipolo momentáneo que se forma en un átomo puede inducir un dipolo complementario en otro átomo. El resultado es una nube electrostática cuyo radio varia y es conocido como **radio de Van der Waals**. Dependiendo de la orientación y distancia de estas nubes, pueden atraerse con mayor o menor fuerza o bien, repelerse si se encuentran demasiado cerca. Aunque la interacción como tal es débil, el número de interacciones es grande en la proteína. Además, no solo son importantes para el plegamiento si no también para reconocimiento en interacciones proteína-proteína y proteína-ligando.

- **Efecto hidrofóbico**: Podemos definir al efecto hidrofóbico como la preferencia de los residuos con cadenas laterales no polares hacia un medio no acuoso. Debido a que hay varios aminoácidos con cadenas laterales hidrofóbicas y no polares, se orientan hacia el interior de la proteína formando un **núcleo hidrofóbico** donde la mayoría de estas cadenas se asocian de manera cercana entre si y lejos del solvente, en este caso, agua. Cabe mencionar que todos los aminoácidos hidrofóbicos o no polares de la proteína se encuentran en este núcleo ya que pudieran estar en la superficie de la proteína para interaccionar con otras cadenas polipeptídicas u otro tipo de sustratos.

Existen dos modelos que tratan de explicar como la estructura terciaria se encuentra codificada dentro de la estructura primaria. El primer modelo es el **modelo local**, donde la interacciones de solo algunos aminoácidos son importantes para el plegamiento (aproximadamente 10 a 20% de los residuos). Por otro lado, tenemos el **modelo global**, en el cual el plegamiento de la proteína depende de la interacción de todos los aminoácidos en la secuencia. Existen evidencias que apoyan ambos modelos, sin embargo el modelo local cuenta con un mayor número de evidencias, ya que se pueden observar proteínas con el

mismo plegamiento y que solo comparten un 10% o menos de los aminoácidos a lo largo de la secuencia [8]. Este hecho implica que unos cuantos aminoácidos se encuentran formando un **núcleo de plegamiento** [9] que es crítico para la estructuración de la proteína y que está conservado en aquellas proteínas que presentan la misma estructura.

Retomando el modelo global, es un hecho que cambios en la secuencia en posiciones que no son parte del núcleo de plegamiento, perturban la cinética de plegamiento de la proteína, sin embargo en la mayoría de los casos no evitan que la proteína adquiera su conformación final [8].

La variabilidad o divergencia que podemos observar entre dos proteínas con el mismo plegamiento, pero con un porcentaje bajo de residuos en común, no es el resultado de cambios azarosos, si no mas bien de cambios correlacionados a lo largo de la proteína, donde para cada posición que varíe, encontramos un cambio compensatorio en otra región de la proteína para mantener la estructura [10, 11].

Relación entre estructura y función

El número de combinaciones posibles que podemos obtener para una proteína de 100 residuos es de 20^{100} secuencias. Es claro que las secuencias biológicas solo observamos una pequeña fracción de todas las permutaciones posibles. El número de secuencias únicas anotadas en las bases de datos de proteínas es de aproximadamente 4 millones. Sin embargo, dentro de esta fracción de secuencias, el número estructuras que a la fecha se han podido observar, es de algunas decenas de miles. Dentro de la fracción de estructuras conocidas, el número de plegamientos únicos que podemos observar es aun menor (algunos cientos) y se estima que posiblemente solo puedan existir unos cuantos miles de plegamientos únicos [12].

Por otro lado, el número de reacciones enzimáticas que se pueden llevar a cabo ya sea observadas, o bien químicamente posibles, sobrepasa el número de plegamientos conocidos. Esto nos puede llevar a suponer que conforme se observen nuevos plegamientos, éstos serán capaces de realizar nuevas funciones enzimáticas; o bien, como se observa actualmente que proteínas con un mismo plegamiento, pueden llevar a cabo diferentes funciones enzimáticas y que por lo tanto, conforme se puedan resolver nuevas estructuras, éstas puedan presentar un plegamiento ya conocido.

Debido a que la secuencia primaria determina tanto la función como la estructura, el compromiso de ambas características está implícito. Si una proteína no confiere una función que pueda permitir al organismo adaptarse a su medio de una mejor manera, esa proteína no persistirá independientemente de que pueda plegarse o no.

Evolución de las proteínas

A lo largo de millones de años, las secuencias de proteínas han variado en los diferentes organismos, sin embargo a pesar de que una secuencia puede variar, no lo hará de tal manera que la función o la estructura terciaria pueda verse alterada de manera drástica. Después de la especiación, dos proteínas que llevan a cabo la misma función han sufrido **mutaciones** de manera independiente pero en ambos casos, la aceptación de las mutaciones están sujetas a restricciones funcionales y estructurales [2].

El tipo de mutaciones que puede sufrir un gene pueden provocar la aparición o desaparición de un residuo, cambiar el tipo de aminoácido en una posición determinada (mutación no-sinónima) o no (mutación sinónima). Cualquiera de estas mutaciones, pueden resultar ventajosas, desventajosas o neutrales para la perpetuación de la proteína según el tipo de presión al cual esté sujeto el organismo que la posee [2].

Cuando existe una **duplicación** del gene que codifica para una secuencia dada, una de las dos copias del gene puede sufrir cambios o mutaciones sin ninguna presión y así dar como resultado una proteína que pueda realizar una nueva función. En el caso de que la nueva función confiera una ventaja al organismo, dicho gene se seleccionará para perpetuarse [2].

El resultado final que podemos observar de dos proteínas que proceden de un mismo origen, es de ligeras modificaciones en la función, reconocimiento de una gama más amplia de sustratos, funciones diferentes y por último, nuevas funciones. Una vez más, podemos mencionar que a pesar de que las modificaciones pudieran ser drásticas, la estructura se mantiene.

Por lo tanto, un gene que codifica para una proteína, con un plegamiento y función en particular, puede dar origen a nuevas funciones, aunque no es la única manera en la cual se puede lograr esto. Sin embargo, este proceso nos explica como las proteínas comparten un origen en común, lo cual nos lleva al concepto de **homología**.

El concepto de homología

Homología es la relación entre dos **caracteres** cualquiera, que descienden de un **ancestro común**, normalmente estos caracteres presentan variabilidad o **divergencia**. Este concepto es de suma importancia en biología, ya que a partir de él se derivan otras definiciones y suposiciones. Los caracteres pueden ser génicos, estructurales o de comportamiento en un organismo. La distinción entre caracteres y conocer los distintos estados que puede presentar un carácter, es de suma importancia para poder concluir si dos caracteres son homólogos o no. Por lo cual es muy importante recordar que la homología reside en los caracteres y no en sus estados [13].

En el caso de las proteínas, cuando dos secuencias presentan cierto porcentaje de aminoácidos en común y estructura similar, se dice que ambas provienen de un ancestro en común y por lo tanto son homólogas. Es muy importante mencionar que la homología es una propiedad absoluta y un error común es que, dado la comparación entre dos secuencias y el porcentaje de identidad que comparten, se refiera a dicho porcentaje como de homología.

Por otro lado, existen otros términos referentes a relaciones entre caracteres y que son frecuentemente usados en la biología [13]:

- **Ortología**: Es la relación entre dos caracteres homólogos que se encuentran en especies distintas.

- **Paralogía**: Es la relación entre dos caracteres homólogos que provienen de una duplicación del gene para dicho carácter y que por lo tanto, ambos caracteres provienen del mismo organismo.

- **Xenología**: Es la relación entre dos caracteres homólogos que implica una transferencia del gene que codifica a uno de ellos, entre dos organismos de especies distintas y que por lo tanto el ancestro en común no es de línea directa.

- **Analogía**: Es la relación entre dos caracteres que no descienden de un mismo ancestro pero son el resultado de **convergencia** hacia un estado muy similar.

Dado que todos los organismos en este planeta estamos relacionados en mayor o menor grado, es de esperarse que encontremos caracteres en común a lo largo de las distintas especies. Resulta interesante observar como cada proteína en este caso, se ha mantenido a través del proceso de especiación a lo largo de millones de años pero también la variación que ha sufrido en el proceso, en muchos casos hace difícil encontrar la relación entre aquellos caracteres que tienen el mismo origen a nivel genético en este caso.

Bioinformática

A grandes rasgos es el uso de **computadoras** para el manejo de la **información biológica**. Si bien no se cuenta con una definición precisa, toda aquella caracterización, recopilación, predicción, búsqueda y análisis de los componentes moleculares que conforman a los seres vivos, utilizando algoritmos computacionales, se puede considerar como parte de la **bioinformática** [14].

En particular, en este trabajo nos enfocaremos al uso y desarrollo de herramientas computacionales para la predicción y análisis de la estructura terciaria de las proteínas, partiendo y haciendo uso solo de su estructura primaria.

El comparar dos secuencias implica alinearlas de tal manera que podamos ver si la secuencia de aminoácidos que cada una presenta es similar o no. Suponiendo que ambas secuencias presentan el mismo número de residuos, una primera aproximación sería ver cuantos aminoácidos son idénticos. Dicho conteo nos dará el porcentaje de identidad entre dos secuencias. Sin embargo, como es de esperarse, en muchos casos las secuencias no presentan la misma cantidad de residuos a pesar de que ambas secuencias estén relacionadas. Al comparar dos secuencias, no solamente podríamos comparar la identidad de cada uno de los residuos si no también, sino la similitud entre ellos dado las características químicas que poseen para realizar interacciones como se ha mencionado previamente [14].

Para secuencia encontrada en un organismo en particular, con una función y estructura determinada, es posible encontrar secuencias homólogas en otros organismos con lo cual podremos construir una familia donde podremos observar los cambios en dichas secuencias como resultado de las mutaciones que han sufrido a lo largo del tiempo y de la especiación. Si analizamos estos los cambios y contabilizamos el tipo y la frecuencia en la que ocurren, podremos evaluar y calificar la intercambiabilidad de un aminoácido por otro. Matemáticamente es posible representar estas frecuencias como una matriz de 20x20 caracteres (aminoácidos en este caso) que representaran la probabilidad de cambio entre uno y otro carácter. A esto se le conoce como **matriz de sustitución**. El cálculo de puntaje en forma logarítmica (**log-odd score**) para matrices de sustitución en secuencias biológicas, es posiblemente el más exitoso ya que a partir del análisis de la frecuencia observada y esperada de cada uno de los caracteres, podemos obtener la probabilidad de que un aminoácido en este caso, sea sustituido por otro o bien, cuales no pueden sustituirlo. Las matrices más populares en cuanto a comparación de secuencia son las BLOSUM [15] y PAM [16]. Con las matrices mencionadas, podemos saber que ciertos aminoácidos son intercambiables con frecuencia en las familias de proteínas que resultan tener un origen en común, con lo cual podemos evaluar no solo la identidad entre dos secuencias, sino también la **similitud** a nivel de la secuencia primaria. Como se puede suponer, tanto el alineamiento de secuencias, como las matrices de sustitución, son herramientas de suma importancia en la biología y el uso de tecnología de cómputo facilita su aplicación de forma masiva.

Como se mencionó, en muchos casos las secuencias a comparar, no presentan la misma cantidad de residuos, lo cual añade complejidad al alineamiento de las secuencias. Por lo tanto, es necesario insertar espacios en blanco en alguna de las dos secuencias para así poder aquellas regiones donde la comparación tenga un valor significativo. El significado biológico de estas posiciones en blanco se debe a que en una de las dos secuencias han aparecido o desaparecido residuos lo cual es resultado de mutaciones en el proceso evolutivo de la proteína. A esto se le conoce como **indel** (abreviatura de las palabras en inglés **insertion/deletion**) y tiene un peso (en contra) al momento de comparar dos secuencias [13, 14]. De todas aquellas posibilidades que tenemos de alinear dos secuencias, la que tomaremos como el mejor alineamiento es aquella donde el **puntaje** sea **mayor**.

Finalmente, un último concepto necesario para este trabajo es el de **perfil**. Un perfil es un modelo evolutivo donde se evalúan la probabilidad de cambios en una familia determinada de proteínas relacionadas, donde la frecuencia de cambios depende de suposiciones *a priori* que se tengan acerca de los caracteres. En el caso de los aminoácidos, se pueden considerar los muchos atributos que presentan para evaluar si en una posición determinada existe un cambio o no, ya que para efectos prácticos, aunque el carácter no sea idéntico, si presenta un mismo atributo no será considerado como diferente. Por lo tanto, el calcular un perfil para una familia determinada, es una mejor descripción de la historia evolutiva de dicha familia [14, 17].

Abstract

Homology detection and protein structure prediction are central themes in bioinformatics. Establishment of relationship between protein sequences or prediction of their structure by sequence comparison methods finds limitations when there is low sequence similarity. Recent works demonstrate that the use of profiles improves homology detection and protein structure prediction. Profiles can be inferred from protein multiple alignments using different approaches. The “Conservatism-of-Conservatism” is an effective profile analysis method to identify structural features between proteins having the same fold but no detectable sequence similarity. The information obtained from protein multiple alignments varies according to the amino acid classification employed to calculate the profile. In this work, we calculated entropy profiles from PSI-BLAST-derived multiple alignments and used different amino acid classifications summarizing almost 500 different attributes. These entropy profiles were converted into pseudocodes which were compared using the FASTA program with an *ad-hoc* matrix. We tested the performance of our method to identify relationships between proteins with similar fold using a non-redundant subset of sequences having less than 40% of identity. We then compared our results using Coverage Versus Error per query curves, to those obtained by methods like PSI-BLAST, COMPASS and HHSEARCH. Our method, named HIP (Homology Identification with Profiles) presented higher accuracy detecting relationships between proteins with the same fold. The use of different amino acid classifications reflecting a large number of amino acid attributes, improved the recognition of distantly related folds. We propose the use of pseudocodes representing profile information as a fast and powerful tool for homology detection, fold assignment and analysis of evolutionary information enclosed in protein profiles.

Resumen

La detección de homología y predicción de estructura de proteínas, son temas centrales dentro de la bioinformática. Establecer relaciones entre secuencias de proteínas o bien, predecir su estructura por métodos de comparación de secuencia, encuentra limitantes cuando la similitud de las secuencias es baja. Trabajos recientes, demuestran que el uso de perfiles mejora los resultados de detección de homología y de predicción de estructura de proteínas. Los perfiles pueden ser construidos, usando diferentes enfoques, a partir de alineamientos múltiples de proteínas. El “Conservacionismo del Conservacionismo” es un método de análisis muy efectivo para identificar características estructurales entre familias de proteínas que presentan la misma estructura pero no similitud a nivel de secuencia. Se calculan valores de entropía para cada posición del alineamiento los cuales son comparados entre las posiciones estructuralmente equivalentes entre las familias. En este caso, la entropía se refiere a un concepto informacional y no a una propiedad termodinámica. La información que se a partir de alineamientos múltiples varía según el tipo de clasificación de aminoácidos que se utilice para calcular el perfil. En este trabajo, calculamos perfiles de entropía a partir de alineamientos múltiples derivados de búsquedas con PSI-BLAST, utilizando diferentes clasificaciones de aminoácidos que resumen casi 500 diferentes propiedades o atributos de los mismos. Estos perfiles de entropía convertidos en pseudocódigos fueron comparados utilizando el programa FASTA con una matriz *ad-hoc.* Probamos el desempeño de nuestro método para detectar relaciones entre proteínas con plegamiento similar usando una base de datos filtrada al 40% de identidad. Comparamos los resultados contra los obtenidos usando otros métodos como PSI-BLAST, COMPASS Y HHSEARCH, empleando curvas de cobertura contra errores por búsqueda. Nuestro método llamado HIP (Homology Identification with Profiles en inglés) presenta gran exactitude detectando proteínas con el mismo plegamiento. El uso de diferentes clasificaciones de aminoácidos que reflejan una gran variedad de atributos, incrementa la capacidad de detección de homólogos remotos con el mismo plegamiento. Hemos propuesto que el uso de pseudocódigos como perfil informacional, puede ser un poderoso método para detectar homología, asignar plegamiento y analizar la información evolutiva codificada dentro de los perfiles.

ANTECEDENTES

Uno de los grandes retos en la bioinformática es la identificación y detección de secuencias homólogas debido a su importancia en la predicción y asignación de estructura terciaria de proteínas [18]. Como ya se ha mencionado, el principio detrás de este enfoque radica en la conservación de la estructura terciaria a pesar de la alta divergencia a nivel de secuencia [19]. Por lo tanto, es posible realizar inferencias acerca de la estructura terciaria de una proteína, solo comparando su secuencia contra las de otras proteínas cuya estructura se conoce. Dependiendo del resultado de dicha comparación, sabremos si las secuencias son homólogas y por lo tanto presentan la misma estructura.

La eficacia de cualquier método de predicción de estructura basado en detección de homología depende de su capacidad de establecer relaciones veraces entre las secuencias. La mayoría de los métodos pierden sensibilidad cuando la relación a nivel de identidad de aminoácidos entre las proteínas es muy poca (< 20%), aunque éstas conserven la misma estructura [20].

Además de poder detectar la relación entre dos secuencias de una manera veraz, se requiere poder encontrar esta relación buscando dentro de bases de datos que contienen secuencias relacionadas con la nuestra y muchas otras que no lo están. **FASTA** es uno de estos programas que puede realizar búsquedas sobre bases de datos ya sea de **ADN** o de proteínas de una manera rápida y precisa. En vez de comparar cada uno de los residuos entre dos secuencias, compara patrones o palabras con lo cual construye un alineamiento local. Gracias a su velocidad y la confiable estadística que usa para evaluar la significancia de sus resultados, FASTA es un método con el cual se puede realizar la búsqueda de secuencias homologas, incluso remotas[21].

Cuando la relación a nivel de secuencia es casi indetectable, el uso de perfiles construidos a partir de alineamientos múltiples puede incrementar la sensibilidad de los métodos para detectar homología remota. Como se mencionó, un perfil es una matriz creada a partir de un alineamiento múltiple de proteínas, donde se asigna un valor de cambio para cada uno de los 20 aminoácidos posibles, dependiendo de su frecuencia en cada posición del alineamiento múltiple. Los perfiles contienen información que nos ayudan a entender factores clave en la evolución de una familia de proteínas.

Varios métodos de comparación de secuencias que emplean perfiles han demostrado ser muy efectivos en la detección de homólogos remotos. Un claro ejemplo es el **PSI-BLAST** [22], el cual es un método que realiza búsquedas iterativas donde en cada búsqueda se construye una matriz de peso (PSSMs) con las secuencias encontradas, para luego emplearse en una nueva búsqueda y así sucesivamente. A pesar de la efectividad de este método, sigue teniendo limites al llegar a la llamada "**zona de penumbra**" (**twilight zone)** donde la relación entre dos secuencias a nivel de identidad de aminoácidos es menor al 20% [23].

Existen otros ejemplos de métodos que emplean comparación de secuencias utilizando perfiles ya sea comparando la secuencia contra el perfil o bien, perfil contra perfil. No está de más decir, que la calidad del alineamiento es de suma importancia para la construcción del perfil.

Los métodos de comparación de perfiles se han convertido en una excelente opción para estudiar las propiedades estructurales en las proteínas, por lo cual dichos métodos son empleados para asignar plegamiento por homología [17].

Perfiles de entropía

En un trabajo previo Mirny y Shakhnovich [24] demuestran que, familias de proteínas con el mismo plegamiento pero sin homología identificable por comparación de secuencia, tienen valores de **entropía** de alta conservación en posiciones estructuralmente equivalentes. Cabe mencionar que entropía en este caso, se refiere a la probabilidad observada de que un aminoácido varíe en una familia de proteínas y no a un concepto termodinámico o energético. Si la conservación de una posición se mantiene entre familias que presentan un mismo plegamiento, dicha posición tiene un papel importante, ya sea para el plegamiento o la función biológica de la proteína, sin importar las características o atributos de los aminoácidos que ahí se encuentren. Este nuevo concepto fue definido por los autores como **Conservacionismo del Conservacionismo (CoC** en inglés).

Los autores evalúan la conservación de cada posición calculando la variabilidad de los aminoácidos en un alineamiento múltiple, conforme a sus propiedades fisicoquímicas: alifáticos, aromáticos, polares, cargados positivamente, cargados negativamente y especiales; se refieren a dichos cálculos como valores de entropía para cada posición. Ellos discuten que en cada familia, las posiciones conservadas son determinantes tanto para el plegamiento ya que las corroboran con datos experimentales que demuestran que dichas posiciones afectan la cinética de plegamientos de las proteínas que toman como referencia, o bien dichas posiciones pertenecen a residuos catalíticos.

Si la entropía de estas posiciones son comparadas contra las de otras familias que presenten el mismo plegamiento, las posiciones equivalentes a nivel de estructura en la otra familia presentan un grado de conservación muy parecido, aunque las propiedades fisicoquímicas de los aminoácidos sean diferentes [9, 24]. Por lo tanto, es posible comparar dichas proteínas independientemente de la similitud que tengan a nivel de aminoácidos.

El cálculo de entropía para una familia de proteínas, utilizando una clasificación de aminoácidos, es un tipo de perfil que nos aporta información de la proteína y de aquellas proteínas homólogas que conforman la familia.

El problema de la métrica en secuencias de proteínas

Recientemente Atchley *et. al.*[25], realizaron un análisis acerca del problema de la métrica en secuencias de aminoácidos, el cual nos habla del problema que existe al usar un alfabeto para caracterizar aminoácidos y encontrar la relación entre las letras de dicho alfabeto. Sin embargo, el orden del alfabeto difícilmente corresponde con las propiedades que cada una de las letras pudiera representar. Estos autores proponen un análisis multidimensional para resolver este problema, donde analizan cerca de 500 atributos que los aminoácidos pueden representar, tales como volumen, tamaño, hidrofobicidad, carga, energía de enlace, ángulos de torsión, etc. [25] Los resultados del análisis estadístico de aquellos atributos que correlacionan unos con otros, quedan resumidos en 5 agrupaciones distintas o factores, donde la distancia entre los aminoácidos esta determinada por las mejores correlaciones entre los atributos evaluados. Dichos factores fueron usados como matrices de sustitución para evaluar la importancia de los residuos en una familia divergente de factores de transcripción y su importancia para el contacto y reconocimiento con su molécula blanco de ADN. Los resultados obtenidos indicaban que ciertos factores funcionaban mejor como modelo evolutivo que varias matrices de sustitución usadas en la comparación de secuencias.

Para construir perfiles a partir de un alineamiento múltiple dado, es posible aplicar distintos criterios según el tipo de propiedad o atributo de los aminoácidos que nos interese observar. Dependiendo de estos criterios, podemos obtener información distinta de la familia de proteínas, como puede ser información estructural, funcional o evolutiva.

Conjuntado estas dos ideas, hemos desarrollado un método de comparación de perfiles de entropía para relacionar familias cuya homología es remota pero que conservan el mismo plegamiento. Este nuevo enfoque fue nombrado como HIP (Homologous Identification with Profiles), el cual emplea perfiles construidos a partir de familias de proteínas homologas. Para cada alineamiento múltiple se calculó la entropía de cada posición utilizando los criterios usados en el trabajo de Mirny y Shakhnovich, así como también hemos implementado las agrupaciones realizadas en el trabajo de Atchley *et al* [25].

HIPÓTESIS

Los perfiles de entropía de familias de proteínas homólogas presentan patrones similares entre si. Por lo tanto, dichos patrones pueden ser comparados y así detectar relaciones entre proteínas homólogas independientemente de la divergencia a nivel de su estructura primaria.

OBJETIVOS

General:

Desarrollar un método de comparación de perfiles de entropía capaz identificar proteínas homólogas remotas, representando dichos perfiles como pseudocódigos para la asignación de plegamiento por homología.

Particulares:

- Calcular los valores de entropía para cada posición de un alineamiento múltiple utilizando distintas agrupaciones de aminoácidos.
- Transformar los valores de entropía a letras de un pseudocódigo.
- Crear un repertorio de pseudocódigos de familias de proteínas homólogas generadas a partir de una secuencia semilla cuyo plegamiento es conocido.
- Implementación de una matriz adecuada para comparar pseudocódigos utilizando algoritmos ya existentes de comparación de secuencias.
- Realizar una comparación entre los resultados obtenidos por nuestro método y aquellos obtenidos empleando otros métodos de comparación secuencia-perfil y perfil-perfil.

METODOLOGÍA

Bases de datos y familias de proteínas homologas

Se empleó la base de datos CATH v2.6 [26, 27] como un set de referencia de dominios de proteínas con un plegamiento ya clasificado. Esta base de datos ofrece una clasificación de plegamientos de forma jerárquica anotada de la siguiente manera (Fig. 6):

- **Clasificación**: Se refiere al tipo de estructuras secundarias que conforman la estructura de una proteína. Pueden ser mayoritariamente alfa, mayoritariamente beta, alfa/beta o pocas estructuras secundarias.

- **Arquitectura**: Se refiere a la disposición en el espacio que observamos de las estructuras secundarias en las estructuras.

- **Topología**: Este nivel de organización subdivide a las estructuras según la conectividad que presentan las estructuras secundarias y que prácticamente define a que tipo de plegamiento corresponden.

- **Homología**: Este nivel jerárquico agrupa a aquellos plegamientos que presentan un origen en común. Se emplean varios criterios para determinar si dos plegamientos son homólogos, como lo son el porcentaje de identidad, el porcentaje de sobrelape entre dominios y valores tanto de alineamiento de secuencia como estructural.

Esta base de datos cuenta con una anotación semi-automatizada, lo cual permite clasificar un mayor número de estructuras. También emplea 5 métodos distintos para encontrar dominios dentro de una cadena, con lo cual se tiene una mejor anotación de los plegamientos por dominio.

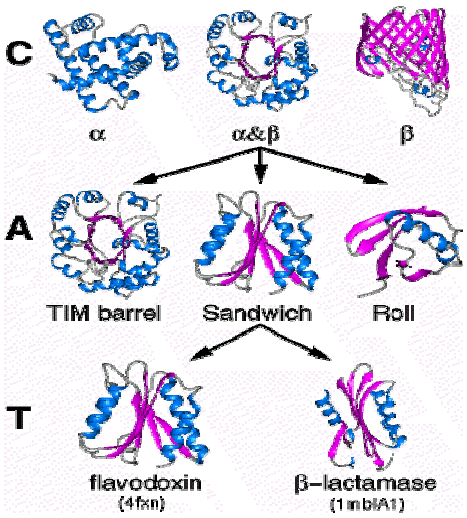


Figura. 6.- Esquema de la jerarquía empleada en la base de datos CATH para clasificar plegamientos. (Modificada de http://www.cathdb.info/latest/index.html)

La base de datos fue filtrada empleando el programa CD-HIT [28], para obtener un grupo de secuencias sin sobrerrepresentación, con un porcentaje de identidad no mayor al 40% entre ellas. El filtrado al 40% nos permite una evaluación sin tener una sobrerrepresentación de proteínas casi idénticas. Los dominios menores a 20 aminoácidos fueron descartados. A este grupo se le denomino CATH40.

Empleando cada dominio del set CATH40, se realizaron búsquedas con el programa PSI-BLAST para la obtención de familias de homólogos. La base de datos utilizada para la búsqueda de proteínas homologas fue la UNIREF90 [29] la cual esta filtrada a un 90% de identidad con un total de 1,388,652 secuencias. Los parámetros utilizados en la búsqueda con PSI-BLAST fueron: Matriz BLOSUM62 para el ciclo inicial, valor de inclusión 0.001 y un máximo de 5 iteraciones y 500 secuencias en el alineamiento. Es importante mencionar que el uso de una base de datos filtrada disminuye la sobrerrepresentación de secuencias casi idénticas y permite incluir secuencias más diversas en el alineamiento final.

A partir de los resultados obtenidos en el último ciclo, se extrajeron las secuencias alineadas con el programa MVIEW [30], con las cuales se calcularon los valores de entropía y pseudocódigos. Para el análisis comparativo, entre HIP y PSI-BLAST, se guardaron las matrices generadas por la última iteración de PSI-BLAST (Position Specific Scoring Matrix, PSSM) para después ser usadas en una búsqueda de un solo ciclo contra CATH40 utilizando cada dominio representante.

Cálculo de entropía y clasificación de aminoácidos:

Se realizó el cálculo de la entropía a partir de los alineamientos múltiples extraídos de la última ronda de PSI-BLAST, empleando la siguiente formula:

$$s(l) = -\sum_{i=1}^{n} p_i(l) \log p_i(l)$$

Donde $p_i(l)$ representa la frecuencia de cada una de las n clases de aminoácidos. Si una posición no presenta aminoácido en por lo menos 30% de las secuencias en el alineamiento, entonces no se calcula el valor de entropía y es considerada como un “gap”.

Se utilizaron 12 clasificaciones de aminoácidos para los cálculos de entropía; 10 clasificaciones fueron obtenidas a partir del trabajo realizado por Atchley *et al*[25] donde se abordó el problema de la métrica en las secuencias de proteínas. Ellos estudiaron 494 diferentes atributos de los aminoácidos, usando un análisis estadístico multifactorial, logrando agrupar todos estos atributos en 5 distintos clasificaciones de aminoácidos representados en dendogramas. De estos 5 dendogramas llamados **factores** del I al V, se realizaron 2 líneas de corte (0.2 y 0.4) sobre la distancia media de las agrupaciones de aminoácidos como se ejemplifica en la Figura 7. También se emplearon las clasificaciones de Mirny & Shakhnovich[24] y la de Lesk, que consisten en 6 agrupaciones de aminoácidos según sus propiedades fisicoquímicas o estructurales respectivamente. La Tabla 1 muestra las 12 clasificaciones utilizadas en nuestro análisis.

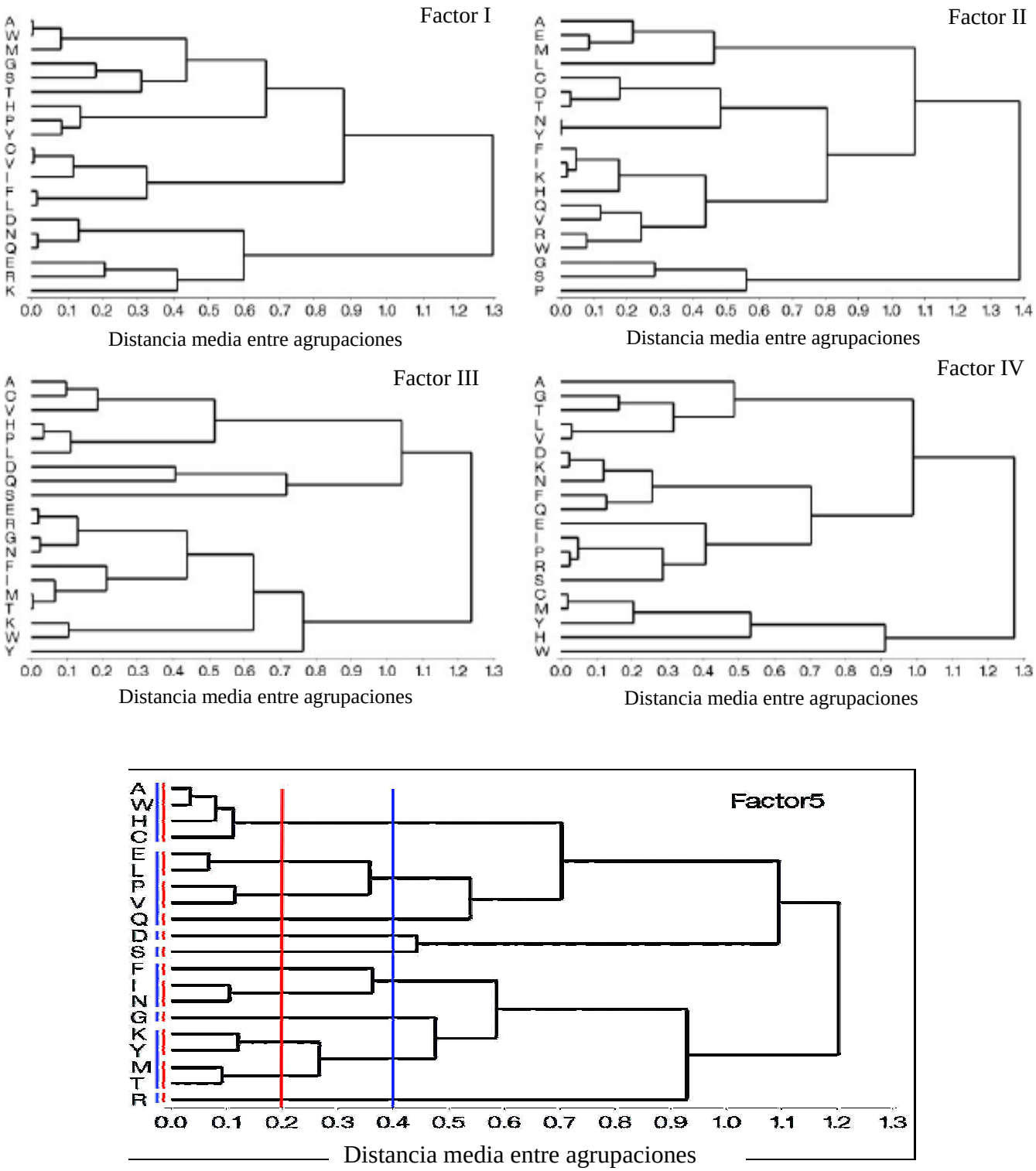


Figura. 7.- Ejemplo de dendogramas representando a los 5 factores. Las líneas de corte según la distancia promedio entre las agrupaciones, están representadas en rojo (0.2) y azul (0.4). Modificada de [25].

Factores	Clasificaciones de aminoácidos											
MS	AVLIMC	WYHF	TQSN	RK	ED	GP						
LESK	AST	CVILWYMPF	HQN	RK	ED	G						
F-Ic2	AWM	GS	HPY	CVI	FL	DNQ	ER	K	T			
F-Ic4	AWM	GST	HPY	CVIFL	DNQ	ER	K					
F-IIc2	A	EM	L	CDT	NY	FIKH	QV	RW	G	S	P	
F-IIc4	A	EM	L	CDT	NY	FIKH	QV	RW	GS	P		
F-IIIc2	ACV	HPL	D	Q	S	ERGN	F	IMT	KW	Y		
F-IIIc4	ACV	HPL	DQ	S	ERGN	F	IMT	KW	Y			
F-IVc2	A	GT	LV	DKN	FQ	E	IPR	S	CMY	H	W	
F-IVc4	A	GT	LV	DKN	FQ	EIPRS	CMY	H	W			
F-Vc2	AWHC	G	LE	KY	IN	PV	MT	R	Q	D	S	F
F-Vc4	AWHC	G	LEPV	KYMT	INF	R	Q	D	S			

Tabla 1.- clasificación de aminoácidos para el calculo de entropía. MS = clasificación fisicoquímica de Mirny & Shakhnovich; LESK = clasificación estructural de Lesk; F-c = Factor-Línea de corte.

Cálculo de pseudocódigos

Una vez calculados los valores de entropía, utilizando las 12 distintas agrupaciones, para cada posición en todas las familias de proteínas en el set de prueba CATH40, se procedió a transformar el perfil de valores de entropía en letras de un pseudocódigo. Dicho código utilizar las mismas 20 letras que los aminoácidos, en este caso cada letra corresponde a un intervalo de valores de entropía.

Para determinar el rango de valores que cada una de las 20 letras representaría, se utilizó el criterio de "Equal Sized Bins" donde los valores mas frecuentes ocupan un rango menor mientras que los valores menos frecuentes se agrupan en un intervalo más amplio. Cada intervalo de valores varía según el número de agrupaciones en la clasificación de aminoácidos.

Matriz de sustitución HIPMAT

Para poder comparar los pseudocódigos entre si, fue necesario crear una matriz adecuada para evaluar la distancia entre los intervalos de entropía. Partiendo de una matriz de identidad, modificamos los valores de tal forma que el valor para la identidad entre dos intervalos es de 4 y decae a 2, 1, 0, -1, -2 y –4 conforme el intervalo comparado se va alejando siendo –4 el valor mas bajo.

Filtrado de pseudocódigos

Para realizar una búsqueda, el pseudocódigo utilizado fue filtrando según su contenido informacional a nivel global [31], secuencias con una entropía de Shannon[31] menor a 2, fueron descartadas. Después del tratamiento mencionado, el número de pseudocódigos que conforman el set CATH40 es de 4,606. El filtrado de los códigos nos ayuda a eliminar regiones de baja complejidad con lo cual se reduce el número de falsos positivos.

Búsquedas con HIP

Utilizamos el programa FASTA (versión 3.3t08d4, Marzo 2001) [21] con los parámetros predeterminados, para comparar nuestros pseudocódigos empleando la matriz previamente descrita. Como se mencionó anteriormente, FASTA es un programa de búsqueda de secuencias con una estadística sólida y una velocidad de procesamiento que permite realizar cientos de comparaciones en cuestión de segundos. A partir del set de pseudocódigos CATH40, seleccionamos 630 pseudocódigos como representantes de cada grupo de secuencias con un mismo plegamiento y fueron filtrados a nivel local utilizando el programa SEG [32] antes de realizar la búsqueda. Cada grupo consiste en al menos dos secuencias del mismo plegamiento. Se realizaron 630 búsquedas por cada uno de los 12 factores contra la base de datos de referencia CATH40 (4,606 dominios) en forma también de pseudocódigo.

Una vez realizadas las 12 búsquedas, se evaluaron los resultados obtenidos de cada búsqueda y para cada caso se tomo el mejor resultado tomando como referencia el valor de expectancia. A esta metodología la llamamos **HIP (Homolgy Identification with Profiles).**

Comparaciones con COMPASS, HHSEARCH y PSI-BLAST

Para evaluar el desempeño de nuestro método, realizamos la misma operación utilizando otros métodos de comparación de perfiles como lo son COMPASS [33] y HHSEARCH [34]. En ambos métodos se utilizaron las mismas familias y representantes para la construcción de perfiles y la búsqueda con las condiciones predeterminadas de cada método. En el caso de HHSEARCH se añadió la información de predicción de estructura secundaria para construir el perfil de cada familia. En el caso del PSI-BLAST, se realizó una sola iteración con los parámetros predeterminados y se empleó cada una de las matrices PSSM generadas en la búsqueda sobre la base de datos UNIREF90.

Curvas CVE

Para evaluar y comparar los resultados obtenidos a partir de las búsquedas con HIP, PSI-BLAST, COMPASS y HHSEARCH realizamos curvas de cobertura *vs* error por búsqueda (Coverage Vs Error per-query [35, 36]). Este tipo de análisis nos permite hacer comparaciones entre métodos cuyos parámetros de evaluación son distintos además de que nos permite evaluar la sensibilidad y especificidad de los métodos. Se tomaron los pares resultantes de las búsquedas y los listamos de mejor a peor resultado según el valor de referencia (valor de expectancia, en este caso) que presenta cada método.

El error por búsqueda es el número total de errores en un punto determinado dividido entre el número total de búsquedas; la cobertura es el número total de resultados positivos en un punto determinado dividido entre el número total de pares de homólogos posibles. Los valores de cobertura y de errores por búsqueda se grafican uno en cada eje y de aquí se puede determinar en que punto se alcanza un porcentaje de error determinado y su correspondiente valor en cobertura, así como también el valor del parámetro de referencia con lo cual se puede obtener una línea de corte.

La cobertura obtenida por cada método incrementa cada vez que un par resulta ser homólogo (CATH) y en el caso contrario se incrementa el número de errores, con excepción en el caso de que los pares obtenidos presentaran los mismos 3 números de notación ya que no existe evidencia suficiente para decir que tienen un mismo origen pero sin embargo presentan una misma topología.

Según la base de datos CATH, si una proteína comparte los mismos 4 números con otra, entonces serán homologas y por lo tanto el par será considerado como un resultado verdadero positivo. En el caso de no compartir la notación, son considerados como error. Si los pares encontrados compartieran solo los 3 primeros números, entonces no serán considerados ni como error ni como resultado verdadero, ya que dichos resultados presentarían la misma topología, esto es, ambas estructuras tienen las mismas estructuras secundarias y están orientadas de manera similar en el espacio, pero no existen mas evidencias para probar que provienen del mismo origen [36].

Cálculo de la prueba de Z

Esta prueba estadística evalúa si los resultados obtenidos por nuestro método presenta algún sesgo en la distribución de los resultados observados y la distribución esperada en el set de referencia, todo con respecto a la clase de plegamiento. Para el cálculo de la prueba (**Z-score**), primero se requiere la desviación estándar (σ) entre la distribución observada (los resultados) y la esperada (CATH40) usando la siguiente formula:

$$\boldsymbol{\sigma} = (X_{obs} - X_{exp})^2 / n$$

Donde la X_{obs} es la distribución de plegamientos de los resultados y la X_{exp} del set de referencia CATH40. Para calcular el Z-score se uso la siguiente formula:

$$\textbf{Z-score} = (X_{obs} - X_{exp}) / \sigma$$

Si el Z-score es mayor a 3 o menor a -3, quiere decir que la diferencia entre los datos observados y esperados es significativa.

RESULTADOS

Evaluación del método

Los resultados de las búsquedas realizadas fueron comparados usando graficas de CVE donde se puede localizar fácilmente a un porcentaje de error, cuanta cobertura tiene un método y viceversa. Además, este tipo de representación nos permite comparar métodos cuyos parámetros y valores de referencia son diferentes [35].

En la Figura 8 se muestra la comparación entre las búsquedas realizadas utilizando PSI-BLAST como se describió en la Metodología y las realizadas con pseudocódigos construidos con las clasificaciones de LESK y la propuesta por Mirny y Shakhnovich[24] (MS). Como se puede observar, las búsquedas realizadas con los pseudocódigos no presentaron un mejor desempeño comparadas con las realizadas con PSI-BLAST.

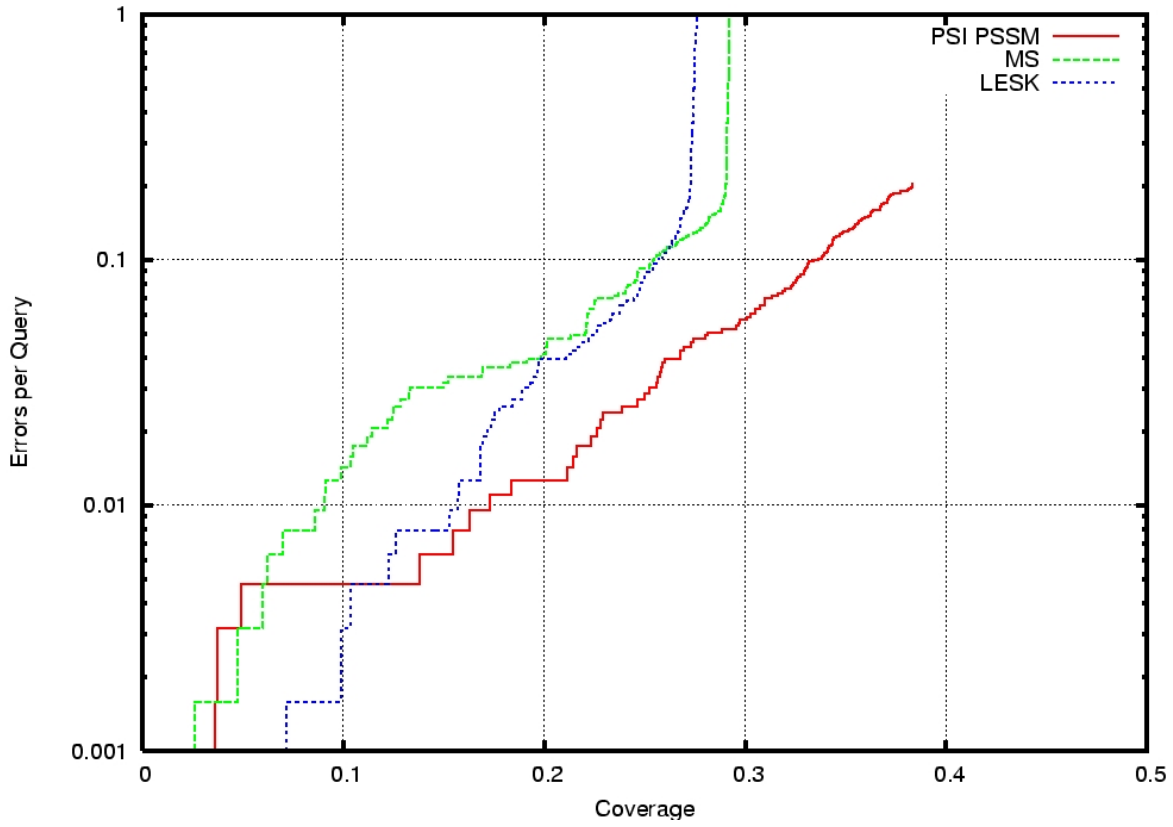


Figura 8.- Gráfica CVE comparando los resultados de las búsquedas usando PSI-BLAST y pseudocódigos. “Error per Query” se refiere al error por búsqueda que se refiere a la fracción de resultados falsos positivos (notación diferente de los 4 números de CATH). “Coverage” se refiere a la cobertura que representa la fracción de verdaderos positivos (mismos 4 números de notación de CATH) con referencia al número total de resultados esperados.

Debido a que los resultados obtenidos con pseudocódigos construidos utilizando las clasificaciones de MS y LESK no presentaron un mejor desempeño, decidimos probar nuevas agrupaciones de aminoácidos para el cálculo de perfiles de entropía, las cuales consideraran otros atributos en los aminoácidos además de las propiedades fisicoquímicas y estructurales. Para ello, utilizamos los factores propuestos por Atchley *et al* utilizando los dendogramas mostrados en la Figura 7 de la Metodología. Con ello, pudimos generara 10 nuevas clasificaciones de agrupación de aminoácidos como se muestra en la Tabla 1. Los resultados de las comparaciones se muestran en la siguiente gráfica (Fig. 9).

Como se puede observar, ninguna de las nuevas clasificaciones mostraron por si solas un mejor desempeño. Sin embargo, el comportamiento de las curvas indica que cada factor obtiene resultados con diferentes valores de referencia. Un mismo resultado podría ser encontrado por varios factores pero con un valor de referencia distinto, así como también un factor en particular podría estar encontrando resultados que con ningún otro factor fuera posible observar.

Por lo tanto, un método que realizara búsquedas simultáneas empleando todos los factores y evaluando cual de ellos reporta el mejor resultado, podría presentar un mejor desempeño que las búsquedas individuales. En la Figura 10 se puede observar la comparación entre los resultados de cada búsqueda por separado y aquella donde se selecciona el mejor resultado de las 12 búsquedas. Es evidente que la selección del mejor resultado proporcionado por cada búsqueda, aumenta significativamente la cobertura del método.

Las búsquedas simultaneas empleando los 12 diferentes factores y la selección del mejor resultado entre ellas, conforman el nuevo método propuesto en este trabajo al cual hemos llamado HIP (Homolgy Identification with Profiles) por sus siglas en inglés. El siguiente paso sería evaluar este nuevo método contra PSI-BLAST y otros métodos empleados para la comparación de perfiles.

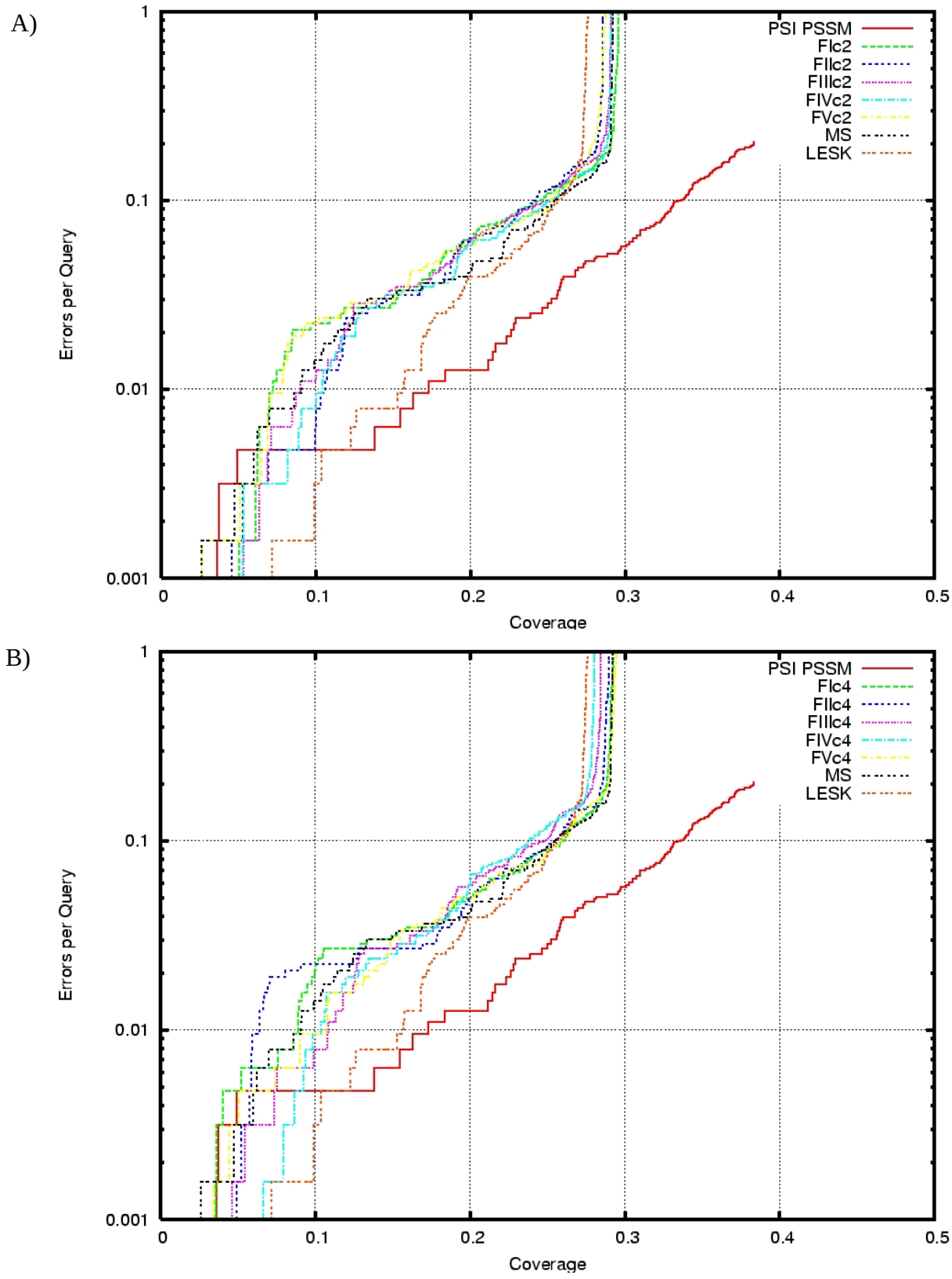
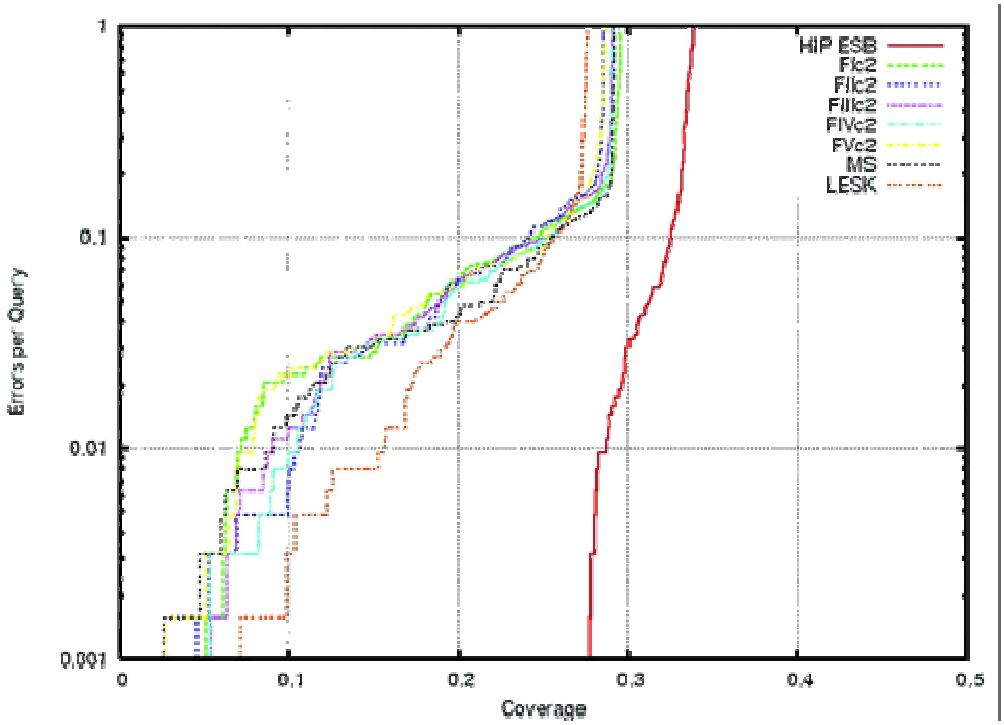


Figura 9.- Gráfica CVE comparando las búsquedas realizadas con PSI-BLAST y pseudocódigos construidos con los difrentes factores. A) búsqueda realizada con cada agrupación a una línea de corte de 0.2. B) búsqueda realizada con cada agrupación a una línea de corte de 0.4.

A)



B)

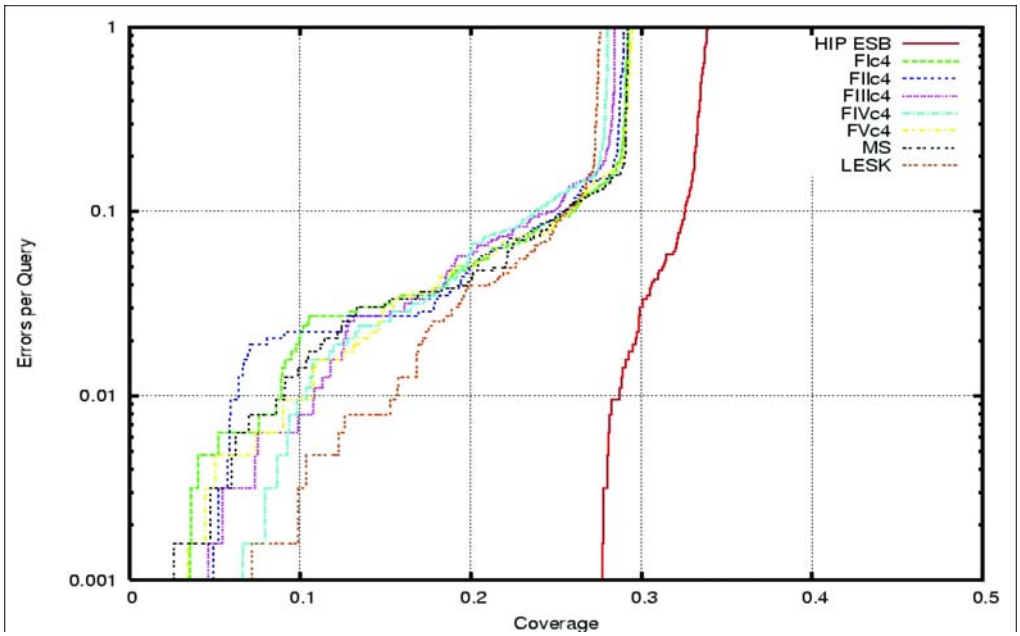


Figura 10.- Curvas de CVE comparando HIP contra búsquedas utilizando los factores por individual. A) búsqueda realizada con cada agrupación a una línea de corte de 0.2. B) búsqueda realizada con cada agrupación a una línea de corte de 0.4.

En la gráfica de la Figura 11, se puede observar como a un porcentaje muy bajo de error (1%) HIP demuestra ser el método con mayor cobertura, sin embargo llegando a un umbral de 10% de error, ya no existe gran diferencia con los otros métodos.

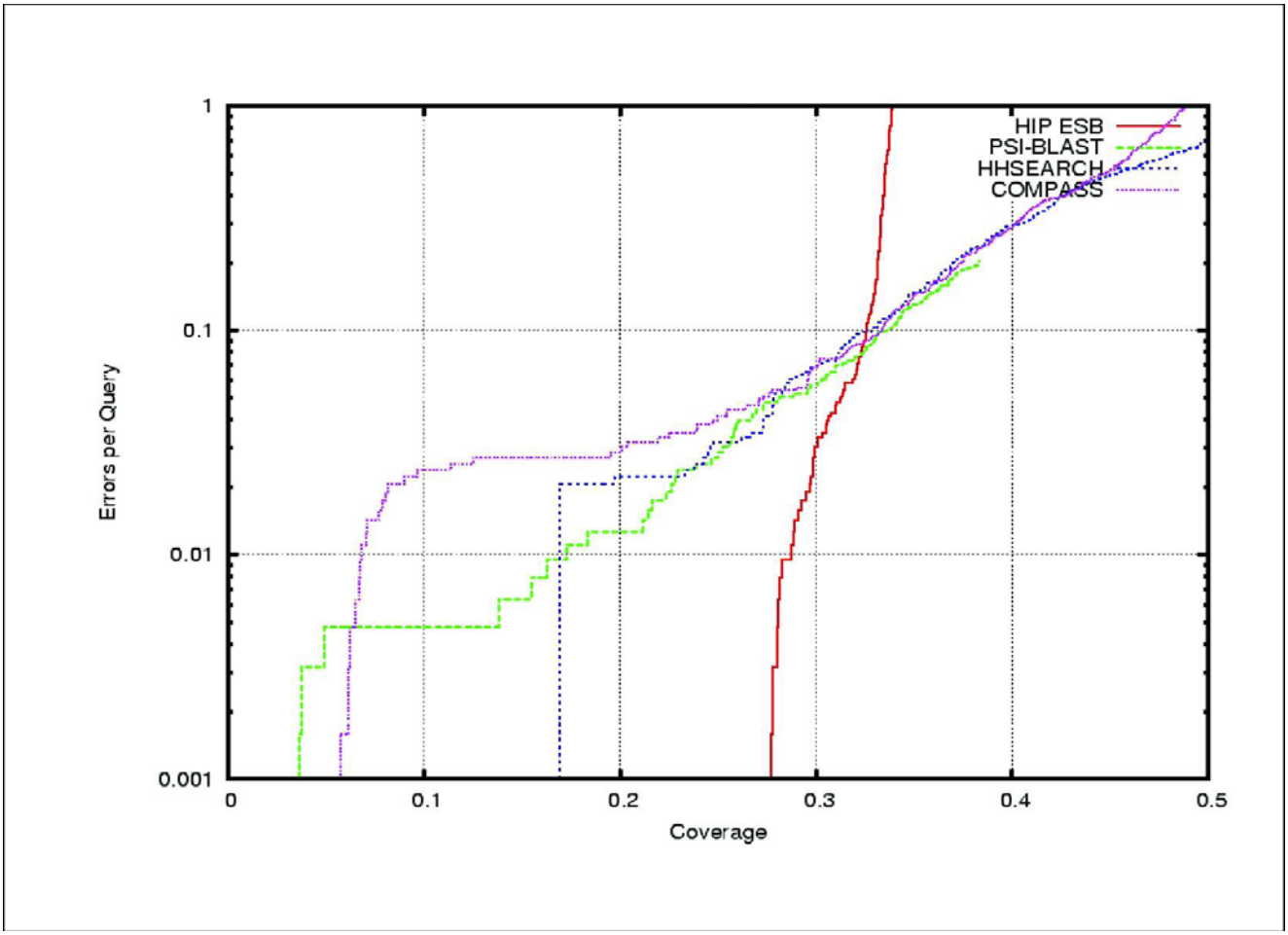


Figura 11.- Gráfica CVE comparando los resultados de los métodos HIP, PSI-BLAST, HHSEARCH y COMPASS. Error per Query se refiere al Error por búsqueda que se refiere a la fracción de resultados falsos positivos (notación diferente de los 4 números de CATH). Coverage se refiere a la Cobertura que representa la fracción de verdaderos positivos (mismos 4 números de notación de CATH) con referencia al número total de resultados esperados.

En la Tabla 2 se muestran los porcentajes de cobertura para el 0%, 1% y 10% de error así como los

valores de expectancia que corresponden a dichos umbrales. Como se observa, nuestro método tiene una alta cobertura sin encontrar ningún error, sin embargo la cobertura no aumenta significativamente aun cuando se permite que el porcentaje de error aumente.

	% Porcentaje de cobertura a un % de error determinado (EPQ)/ E-value		
METODO	*0 % EPQ*	*1 % EPQ*	*10% EPQ*
HIP-ESB	27.70% / 7.4E-13	28.74% / 4.1E-11	32.54% / 0.0011
PSI-BLAST	3.63% / 8E-56	17.30% / 3E-23	33.41% / 1E-07
COMPASS	5.77% / 2.55E-159	6.81% / 2.12E-150	33.31% / 4.14E-37
HHSEARCH	16.97% / 0	16.97% / 0	32.83% / 3E-24

Tabla 2.- Cobertura alcanzada por cada método y su correspondiente porcentaje de error y E-value.

Contribución de cada clasificación de aminoácidos a la selectividad:

HIP presenta una gran selectividad con respecto a los otros métodos empleados. La selección de del mejor resultado para cada búsqueda usando los 12 diferentes factores, incremento el desempeño de nuestro método. Analizamos la distribución de resultados verdaderos positivos para cada factor en el rango de 0% de EPQ para ver la aportación de cada factor. La clasificación del Factor I fue la que contribuyó con la mayor parte de los resultados verdaderos positivos, seguido de la clasificación MS y la del Factor V (Fig. 12). Realizamos el mismo análisis pero tomando en cuenta las líneas de corte de distancia media para los factores y aun así, la distribución no fue afectada. Sin embargo, observamos que para el Factor I y V, la línea de corte de 0.4 presenta una frecuencia particularmente alta de resultados verdaderos positivos (Fig. 13).

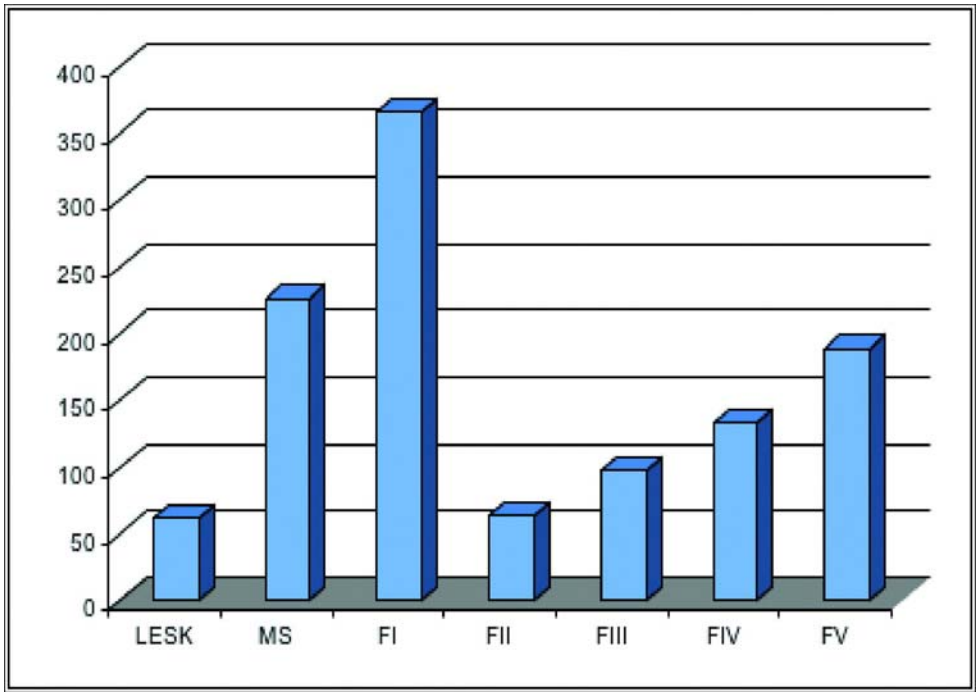


Figura 12.-Distribución de resultados por clasificación. Eje Y =Frecuencia de resultados a 0EPQ. Eje X = Agrupación por los 5 factores.

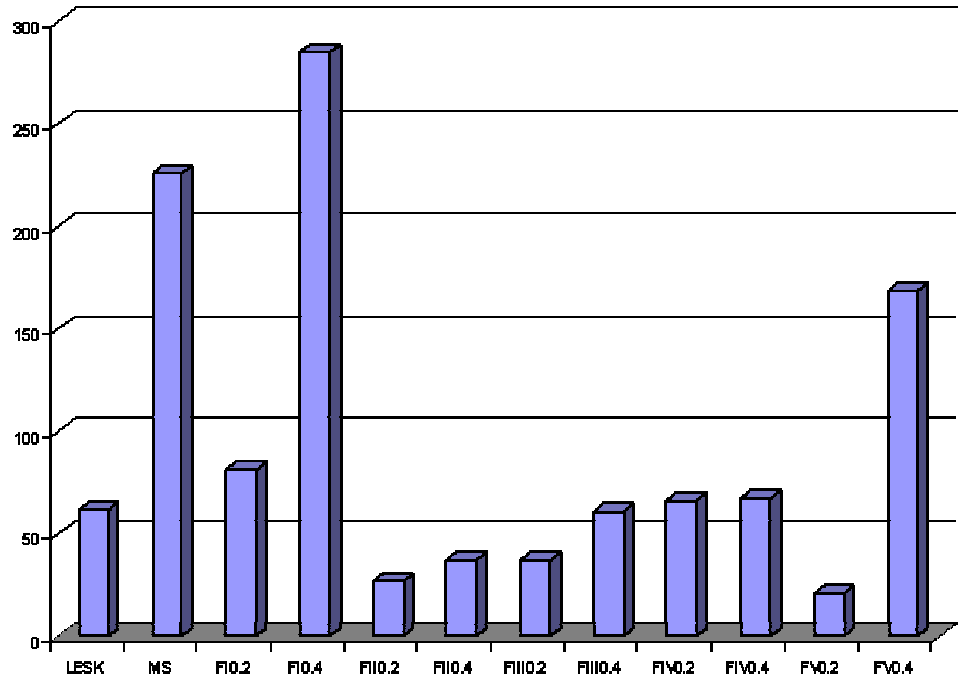


Figura 13.- Distribución de resultados por clasificación y línea de corte. Eje Y = Frecuencia de resultados a 0EPQ.
Eje X = Agrupación por los 12 diferentes factores.

Detección de homólogos remotos a un tasa de 0 EPQ:

En la Tabla 3 se muestran los resultados obtenidos por cada método, divididos por rangos de identidad de aminoácidos de cada uno de los pares encontrados a una tasa de error de 0 EPQ. HIP obtiene el mayor número de resultados para todos los intervalos. Resulta interesante que nuestro método reporta una cantidad considerable de resultados en los intervalos de baja identidad (menos de 20%).

Rango de %ID *(comparación a nivel de aminoácidos)*	*NÚMERO DE RESULTADOS*			
	HIP	*COMPASS*	*HHSEARCH*	*PSIBLAST*
X ≤ 10%	**32**	1	21	0
10% < X ≤ 20%	**369**	58	254	11
20% < X ≤ 30%	**535**	125	398	75
30% < X ≤ 40%	**181**	50	137	55
Total	***1117***	*234*	*810*	*141*

Tabla 3.- distribución de resultados por rangos de porcentaje de identidad entre los pares de proteínas encontradas por cada método a una tasa de 0 EPQ.

Se evaluaron los resultados obtenidos por HIP según su distribución por clase de plegamiento (clasificación de CATH), con la finalidad de ver si nuestro método presenta algún sesgo en encontrar una clase de plegamiento en particular (Tabla 4).

Rango de %ID *(comparación a nivel de aminoácidos)*	# de resultados	May. Alfa	May. Beta	Alpha/Beta	Pocas Est. Sec.
X ≤ 10%	32	15	7	10	0
10% < X ≤ 20%	369	91	90	185	3
20% < X ≤ 30%	535	129	111	285	10
30% < X ≤ 40%	181	40	30	97	14
Total (%)	***1117 (100%)***	***275 (24.62%)***	***238 (21.31%)***	***577 (51.65%)***	***27 (2.42%)***

Tabla 4.- distribución de clase de plegamientos y rangos porcentajes de identidad de resultados a una tasa de 0 EPQ. May. = Mayoritariamente; Est. Sec. = Estructuras Secundarias.

Para saber si las diferencias encontradas en la Tabla 4 son significativas, calculamos el Z-score para la distribución de los resultados con respecto a la distribución que tiene nuestro set de referencia CATH40. Un Z-score mayor a 3 o menor a -3 indicaría que la distribución de los resultados es diferente de la distribución del set de referencia, lo cual indicaría un sesgo en la detección de alguna clase. Sin embargo, ninguna de las diferencias fue significativa como se muestra en la Tabla 5.

Clase (notación CATH)	Distribución de clase de plegamiento		Z-score
	En CATH40	En resultados 0 EPQ	
May. Alfa	21.24%	24.62%	1.46
May. Beta	23.05%	21.31%	-0.78
Alpha/Beta	54.03%	51.65%	-1.06
Pocas Est. Sec.	1.58%	2.42%	0.37

Tabla 5.- Significancia de la diferencia entre la distribución de clases de plegamiento de los resultados obtenidos por HIP y el set CATH40. May. = Mayoritariamente; Est. Sec. = Estructuras Secundarias.

De los resultados obtenidos, HIP fue capaz de encontrar 14 nuevas relaciones (Tabla 6) no encontradas por los otros métodos, bajo las condiciones y parámetros de búsqueda descritos. De estos resultados nos parece particularmente interesante la relación entre los dominios 1b73A1 y 1fjlA2, que corresponden a la glutamato racemasa de *Aquifex pyrophilus* y la aspartato racemasa de *Pyrococcus horikoshii*, respectivamente.

Dichas enzimas comparten un 13% de identidad global a nivel de aminoácidos. Reportes previos [37, 38] sugieren que estas dos estructuras pudieron haber provenido de un ancestro común, debido a la conservación de residuos en el sitio activo y la similitud de sus elementos estructurales. Los resultados de la búsqueda contra CATH40 usando ambos dominios como semilla, se presentan en la Figura 14. En ambas búsquedas los resultados son similares, variando solamente el factor que reporta el mejor resultado. La región de sobrelape en ambos dominios corresponde a una misma región estructural y es interesante como pseudocódigos calculados con diferentes clasificaciones pueden describir una misma región.

Semilla	Resultado	% Ide. Aminoácidos
1lyaB0	1w50A1	5.00%
1jf0A0	1ggwA1	8.10%
1b73A1	1jflA2	12.20%
1m3vA0	1g47A0	13.40%
1f4sP0	1pyc00	15.00%
1f4sP0	1aw600	15.90%
1md8A3	1gknA1	18.10%
1lr7A0	1r0tB0	22.50%
1lr7A0	1tbrR2	25.70%
1tbn00	1faq00	25.80%
1jrfA0	1f5yA1	34.60%
1jrfA0	1cr8A0	35.40%
1jrfA0	1d2lA0	38.00%
1jrfA0	1f5yA2	44.70%*

Tabla 6.- Pares obtenidos únicamente por HIP y el porcentaje de identidad a nivel de aminoácidos al ser alineados los dos dominios. * El porcentaje de identidad es el resultado de un alineamiento global en condiciones predeterminadas, por lo que el porcentaje de identidad en este caso, supera valor de filtro de la base de datos (40%) ya que el alineamiento por pares que se utiliza para filtrar la base de datos es diferente.

Search result of 1jflA2 vs CATH40 pseudocode database

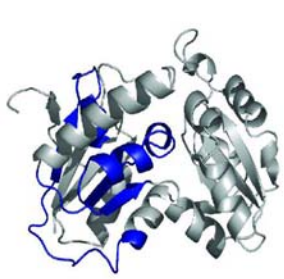
```
FASTA searches a protein or DNA sequence data bank version 3.3t08d4 Mar. 27, 2001
The best scores are:                                          opt bits E(5354)
1jflA2 479 3.40.50.1860                              ( 106)  245  155 1.5e-39
1b73A2 461 3.40.50.1860                              ( 106)  119   79 1.3e-16
1b73A1 455 3.40.50.1860                              ( 100)   94   64 4.2e-12


                                        10        20        30        40        50        60        70
1jflA2                                 TKHVVNSRRVDHADHAYQCHWPWSIYWVEYWSNQPLPKMSTTQYWSMQTVFLTRGNWHSN***********
                                        .......:::::::.:::::::.::::::..:..
1b73A1 HFAAFIPNIHSYENTYMTEECCLGKNPLVYLKYWPWSRVDHADGAYQCHWNWSIYWTFYYRSVTMVDSVP
               40        50        60        70        80        90       100
```

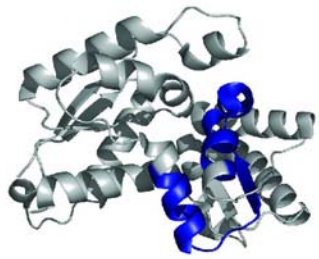
Search result of 1b73A1 vs CATH40 pseudocode database

```
FASTA searches a protein or DNA sequence data bank version 3.3t08d4 Mar. 27, 2001
The best scores are:                                          opt bits E(5353)
1b73A1 455 3.40.50.1860                              ( 100)  208  130 3.9e-32
1b73A2 461 3.40.50.1860                              ( 106)   89   60 7.2e-11
1jflA2 479 3.40.50.1860                              ( 106)   66   46 9.2e-07


         70        80        90       100       110       120       130
1b73A1 CFCAAEGIKHGVYCKWWLWFFEELFIKMFPTGGTQITNQSCIADIAYPCDSMYPIYWNEYTPHNNRNDRLEHQ
                                    ....:......::::.:.....:..::.....
1jflA2                             ILVKERQMSPRQDIADICYSDEVPYNGYWRFWWRRQQLNPPVRWPWWSIPVSIKTQFPYIYPYWPTWPHWYN
                                        10        20        30        40        50        60        70
```



1jflA



1b73A

Figura 14.- Resultados de HIP para las búsquedas usando los dominios 1jflA2 y 1b73A1. Se muestran los alineamientos de pseudocódigos y los valores de expectancia de FASTA. Los resultados bidireccionales en cada tabla aparecen en negritas. Las regiones alineadas que sobrelapan están coloreadas en azul, tanto en el alineamiento como en la estructura.

DISCUSSION

Hemos desarrollado un nuevo método para detectar homología y asignar plegamiento, usando solo la información contenida en los alineamientos múltiples. Nuestro método emplea la idea del Conservacionismo del Conservacionismo propuesta por Mirny & Shakhnovich [24] donde evaluaron estadísticamente el grado de conservación en las posiciones de un alineamiento. Ellos observaron la correlación entre los valores muy conservados dentro de una familia y en las posiciones equivalentes estructuralmente, en otras familias con que presentaban el mismo plegamiento pero no relacionadas a nivel de secuencia. Las posiciones muy conservadas que mantenían el mismo nivel de conservación en todas las familias fueron relacionadas con datos experimentales y resultaron ser posiciones claves para el plegamiento.

Con este enfoque, decidimos transformar los valores de entropía a una representación de pseudocódigo empleando las mismas letras que en el alfabeto de aminoácidos. Con un pseudocódigo con el mismo alfabeto usado para los aminoácidos, es posible hacer uso de algoritmos (FASTA en este caso) destinados para comparación de secuencias proteicas.

Para este fin, desarrollamos una matriz que pudiera evaluar la distancia entre los caracteres (que representan rangos de entropía) en los pseudocódigos. Elegimos partir de una matriz de identidad donde el valor de la diagonal es de 4 y cualquier otra relación esta penalizada con -4. Estos valores vienen de los cálculos para matrices tipo “log-odd”, partiendo de dos secuencias idénticas. No sabemos *a priori* la frecuencia de cambios entre las letras de nuestro código, sin embargo, esperábamos que el perfil de entropía descrito por los pseudocódigos se conservara entre las familias que presentan un mismo plegamiento y que solo se desplazara el rango de valores entre ellos dependiendo de su divergencia. Otro supuesto fue que valores de entropía muy alejados no tendrían porque tener relación. Con estos razonamientos, construimos la matriz HIPMAT, la cual nos permitió encontrar familias de proteínas que presentan el mismo plegamiento y sus perfiles de entropía son similares y no solamente a nivel de las posiciones más conservadas. Una de las grandes ventajas de nuestro método es que al usar un código cuyas letras son las mismas que los aminoácidos, el uso del programa FASTA nos permite hacer miles de búsquedas en cuestión de minutos.

Comparamos el desempeño de nuestro método contra otros métodos de comparación secuencia-perfil (PSI-BLAST) o perfil-perfil (COMPASS y HHSEARCH). HIP presento una mejor cobertura a niveles de error muy bajos aunque, el desempeño de todos los métodos es prácticamente el mismo a una tasa de error de 0.1 EPQ (~10% de errores). Los resultados obtenidos son congruentes con el hecho bien conocido de que los métodos de comparación perfil-perfil son más eficientes para encontrar relaciones de homología en las secuencias de proteínas en la mayoría de los casos [17]. HIP alcanza su límite de cobertura a una tasa de 0.1 EPQ, esto se puede explicar fácilmente ya que un método que funciona con gran especificidad se ve afectado en la sensibilidad [36]. Sin embargo, la cobertura alcanzada por el método, sin reportar ningún error, es excelente. Dichas características son deseables en un método predictivo automatizado.

Analizamos los resultados obtenidos a una tasa de 0 EPQ pasa así poder evaluar el número de relaciones que al ser comparadas usando su secuencia de aminoácidos, presentaran porcentajes de identidad dentro de la zona de homología remota (menos del 20% de identidad). Como se muestra en la Tabla 3, 401 resultados de 1,117 presentan identidades de 20% o menos, demostrando que HIP es capaz de detectar un número considerable de relaciones con identidades remotas. Esto nos sugiere que los perfiles de entropía pueden ser utilizados para detectar proteínas con plegamiento similar sin utilizar ninguna información estructural. Los reportes anteriores [17] nos demuestran con un análisis estadístico detallado y con evidencias de cinéticas de plegamiento, que las posiciones con altos valores de CoC están relacionadas con residuos catalíticos o de importancia estructural.

Uno de los resultados detectados solo por nuestro método, es la relación entre el dominio 1b73A1 de la glutamato racemasa y el dominio 1jflA2 de la aspartato racemasa. Estos dominios comparten alrededor de un 13% de identidad a nivel de aminoácidos. Encontramos interesante esta relación debido a su bajo porcentaje de identidad y la distancia filogenética de los organismos a los que pertenecen (*Aquifex pyrophilus* [Bacteria] y *Pyrococcus horikoshii* [Arquea]). Existen trabajos previos acerca de estas enzimas donde se propone que ambas tienen un ancestro en común y un mecanismo evolutivo muy similar [37-40]. El mecanismo evolutivo que proponen los autores para la evolución de la aspartato racemasa, involucra una duplicación de un gen ancestral, seguido de una fusión y adaptación de los dominios para dar origen a la enzima con una simetría especular capaz de reconocer a ambos enantiómeros del ácido aspártico. La posición que sobrelapa de la aspartato racemasa sobre la glutamato racemasa corresponde a la hélice $\alpha 7$. Dicha estructura esta relacionada con la teoría evolutiva propuesta por los autores y es la que superimponen estructuralmente con el otro dominio dentro de la aspartato racemasa. En la Figura 14 se puede apreciar como la región de la aspartato racemasa sobrelapa en dos zonas contiguas de la glutamato racemasa, dependiendo del factor con el que fue construido el perfil de entropía. Según evidencias previas[37], debido a que la glutamato racemasa cuenta con 40 aminoácidos extras en el dominio relacionado, la estructura se distorsiona y la sobreposición estructural no demuestra con claridad la relación. Sin embargo, otras referencias como la base de datos SCOP [41] reportan a estas dos enzimas como homologas y miembros de una misma familia estructural. Esto corrobora que la comparación de perfiles de entropía es una alternativa confiable que no depende de información ni alineamiento estructural ni de identidad en la secuencia, lo que sugiere que el patrón de perfil se conserva como una especie de “firma del plegamiento”.

En el análisis de CoC, los valores de entropía son calculados usando un esquema de clasificación de aminoácidos por sus propiedades fisicoquímicas y solo las posiciones muy conservadas y estadísticamente significativas, fueron justificadas con evidencia experimental. Sin embargo, en nuestra evaluación el uso de esta clasificación no obtuvo la mayoría de los mejores resultados, además de que en los pseudocódigos alineados tanto las posiciones conservadas como las variables, son reconocidas dentro de la búsqueda. En nuestro análisis, el Factor I fue el que obtuvo la mayor frecuencia de los resultados esperados (Fig. 12). De acuerdo con los resultados encontrados por Atchley *et al.*, esta agrupación tiene correlación con un gran número de matrices de sustitución, lo que sugiere que existe una base evolutiva sólida asociada a los múltiples atributos descritos por esta agrupación. En particular, las agrupaciones de aminoácidos obtenidas con la línea de corte de 0.4 (Tabla 1), presenta una alta frecuencia de resultados con el mejor valor (Fig. 13). Ninguna búsqueda por si sola muestra un mejor desempeño comparada con HIP (Fig. 10).

El uso de los 12 diferentes esquemas de clasificación de aminoácidos y la selección del mejor valor de los resultados de las 12 búsquedas simultaneas, incremento significativamente la especificidad de nuestro método. Es interesante que ciertas relaciones solamente son encontradas utilizando ciertas clasificaciones de aminoácidos. En la mayoría de los casos, existe un consenso de las 12 búsquedas en cuanto a los resultados reportados, sin embargo cuando la relación es más distante solo algunas búsquedas reportan un resultado verdadero positivo. Esto nos sugiere que la historia evolutiva de una familia de proteínas, no puede ser explicada usando un solo esquema de clasificación de aminoácidos y que son muchos los atributos que se pueden evaluar para encontrar la relación entre dos proteínas a pesar de la divergencia en la secuencia.

Adicionalmente, revisamos que nuestro método no presentara sesgo hacia detectar alguna clase de plegamiento en particular. Calculamos el Z-score para saber si las diferencias entre las frecuencias observadas y las esperadas, eran significativas. Como se puede ver en la Tabla 5, ninguna de las distribuciones presenta variaciones significativas, lo que nos indica que los resultados siguen la misma distribución presente en la base de datos.

Existen ciertas consideraciones acerca del método que son importantes mencionar, ya que afectan el desempeño de un método de comparación de perfil-perfil: la calidad del alineamiento; el nivel de divergencia de las secuencias en el alineamiento; el contenido de "gaps" y la complementación con información estructural o evolutiva. Cuando un alineamiento múltiple contiene muy poca información debido a que contiene pocas secuencias o un gran contenido de "gaps" o muy poca divergencia entre las secuencias, no es posible generar pseudocódigos informativos. Para evitar esta complicación, analizamos el contenido informacional global de los pseudocódigos calculados usando la teoría informacional de Shannon [31], dejando fuera del análisis todas aquellas secuencias que tuvieran un contenido informacional menos a 2. Previo a la búsqueda, en cada semilla enmascaramos aquellas regiones de baja complejidad usando el programa SEG [32]. Estos filtros nos ayudaron a disminuir la cantidad de resultados falsos positivos.

HIP no ofrece buenos resultados si se analizan perfiles de entropía que provengan de alineamientos con menos de 20 secuencias, así como tampoco alineamientos donde todas las secuencias se encuentran muy conservadas y el grado de divergencia es muy bajo. Esto se debe a que el pseudocódigo generado no será informativo para el análisis. Sin embargo, pensamos que debido a la tendencia exponencial con que crecen las bases de datos de secuencias, estas limitantes desaparecerán.

CONCLUSIONES Y PERSPECTIVAS

Nuestro método es capaz de relacionar proteínas con el mismo plegamiento con gran exactitud. Los resultados demuestran que el uso de pseudocódigos, que representan información evolutiva y estructural encerrada en los alineamientos múltiples, puede ser una herramienta rápida y poderosa para detectar homología y asignar plegamientos. Además, el análisis del alineamiento de pseudocódigos así como refinamiento, es necesario para entender a profundidad las firmas descritas por los perfiles de entropía y el significado biológico que se encuentran representando. Sabemos de la importancia de las posiciones conservadas dentro de los perfiles, sin embargo aun no tenemos claro la relación de las posiciones no conservadas pero que se encuentran formando parte del perfil y que al parecer, la variabilidad en las familias de proteínas, no es completamente azarosa.

HIP puede fácilmente implementarse como un método automatizado para la predicción de estructura debido a su gran especificidad ya que no requeriría de supervisión constante de los resultados. El uso de búsquedas simultaneas usando las 12 diferentes clasificaciones de aminoácidos y la selección del mejor resultado, mejora significativamente el desempeño del método para reconocer relaciones de homóloga distantes y plegamientos. Hemos explorado los distintos atributos que un amino ácido puede tener, de una manera práctica y sencilla lo cual también contribuye a la obtención de resultados que contienen diferente información.

Estudios acerca de la correlación entre que familias son reconocidas por cuales agrupaciones, podría ayudar a entender las propiedades especificas que describen la relación entre las proteínas que comparten cierto plegamiento y por lo tanto su historia evolutiva. Es posible que las fuerzas evolutivas que dirigen el plegamiento de una proteína o su estabilidad, dependan no solo de las propiedades estructurales de los aminoácidos. Se requiere de un análisis mas profundo de las regiones alineadas en las búsquedas de pseudocódigos para alcanzar las metas antes mencionada. Además, el mejoramiento de la matriz de distancia es necesario y se puede lograr evaluando las frecuencias de cambios en los pseudocódigos alienados usando un cálculo del tipo “log-odd”, como son calculadas muchas de las matrices de aminoácidos que se utilizan con mayor frecuencia. En cuanto al tipo de algoritmos que se pueden emplear en vez de FASTA, sería la implementación del método utilizando el programa BLAST [22], el cual podría aumentar la sensibilidad del método con un ligero aumento en el tiempo de cómputo.

En teoría, cualquier alineamiento que cumpla con los requisitos descritos en nuestra metodología, puede ser utilizado para calcular un perfil de entropía y ser comparados contra otros pseudocódigos independientemente del método con que se haya realizado el alineamiento múltiple. Las bases de datos, tanto las de secuencias como las de estructuras, se van enriqueciendo conforme pasa el tiempo y los alineamientos que se pueden generar aumentaran tanto en número como en divergencia. El proceso de actualización de nuestras bases de datos de pseudocódigos, así como también la creación de nuevas bases de datos para realizar estudios de asignación de plegamiento y detección de homologa remota, es un trabajo que se debe de realizar constantemente.

GLOSARIO

(Por orden de aparición)

PROTEINA.- Biopolímero formado por una o varias cadenas de aminoácidos, fundamental en la constitución y funcionamiento de la materia viva.

AMINOACIDO.- Molécula orgánica formada por un carbono quiral, un grupo amino, un grupo carboxilo y un grupo variable que define su propiedad química. 20 diferentes aminoácidos son los componentes fundamentales de las proteínas.

ESTRUCTURA PRIMARIA.- En las proteínas, es la secuencia de aminoácidos que forman el polímero.

ESTRUCTURA SECUNDARIA.- En las proteínas, es la formación de estructuras locales que describen formas tales como hélices, láminas, vueltas o giros.

ESTRUCTURA SUPERSECUNDARIA.- Son motivos de estructuras secundarias que se observan frecuente mente entre proteínas y dentro de las mimas.

ESTRUCTURA TERCIARIA.- Es el arreglo en el espacio de la proteína y que frecuentemente adopta una forma globular.

ESTRUCTURA CUATERNARIA.- Es la interacción entre dominios o cadenas de proteínas.

SUBUNIDAD.- Se refiere a un solo polímero o una sola cadena de una proteína que cuenta con más de una.

PLEGAMIENTO.- Es la descripción detallada del arreglo espacial de la cadena polipeptídica.

ENERGIA LIBRE DE GIBBS.- En termodinámica, es una función de estado, con unidades de energía que marca la condición de equilibrio y la espontaneidad de una reacción química.

ESTADO DE TRANSICION.- Es un estado de alta energía por el cual pasa una proteína antes de adoptar su conformación final, que será un estado de mínima energía.

FRAMEWORK.- Término en inglés que se aplica en proteínas para indicar la formación ordenada de estructuras secundarias que dan como resultado la formación de la estructura terciaria.

DOMINIO ESTRUCTURAL.- Unidad mínima en las proteínas que es capaz plegarse de manera independiente.

MOSAICO.- Proteína formada por varios dominios estructurales con plegamientos distintos.

TOPOLOGÍA.- Descripción detallada del arreglo espacial formado por las estructuras secundarias y su conectividad en el plegamiento de una proteína.

UP-AND-DOWN.- Término en inglés que se refiere a una estructura supersecundaria y que describe un

patrón de dos hélices conectadas que suben y bajan en un mismo eje.

MEANDER.- Término en inglés que se refiere a una estructura supersecundaria y que describe un patrón de dos hojas beta que describen una forma similar al meandro de un río.

GREEK-KEY.- Término en ingles que se refiere a la estructura supersecundaria formada por la repetición de meandros y que describe una forma muy similar a la llave griega en arquitectura.

RANDOM COIL.- Término en inglés que se refiere a una estructura secundaria que no es ni hélice alfa ni hoja beta.

BACKBONE.- Término en inglés que se refiere al esqueleto de carbono de la proteína.

ADN.- Ácido DesoxiriboNucleico.

ANEXO

PROTEINS:
Structure, Function, and Bioinformatics

Protein homology detection and fold inference through multiple alignment entropy profiles

Journal:	*PROTEINS: Structure, Function, and Bioinformatics*
Manuscript ID:	Prot-00721-2006.R1
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	21-Feb-2007
Complete List of Authors:	Segovia, Lorenzo; Instituto de Biotecnología UNAM, Departamento de Ingeniería Celular y Biocatálisis SÃinchez-Flores, Alejandro; Instituto de Biotecnología UNAM, Departamento de Ingeniería Celular y Biocatálisis Pï¿½rez-Rueda, Ernesto; Instituto de Biotecnología UNAM, Departamento de Ingeniería Celular y Biocatálisis
Key Words:	Fold recognition, homology detection, entropy profiles

Title:

Protein homology detection and fold inference through multiple alignment entropy profiles

Short title:

Finding Protein homology with entropy profiles

Alejandro Sánchez-Flores, Ernesto Pérez-Rueda and Lorenzo Segovia[§]
Departamento de Ingeniería Celular y Biocatálisis, Instituto de Biotecnología, Universidad Nacional Autónoma de México.

[§]Corresponding author
Lorenzo Segovia
Departamento de Ingeniería Celular y Biocatálisis Instituto de Biotecnología, Universidad Nacional Autónoma de México.
Apdo. Post 510-3Cuernavaca, Morelos, 62250, México.
Tel +52(777) 3291862
Fax +52(777) 3172388
Email: lorenzo@ibt.unam.mx

Abstract

Homology detection and protein structure prediction are central themes in bioinformatics. Establishment of relationship between protein sequences or prediction of their structure by sequence comparison methods finds limitations when there is low sequence similarity. Recent works demonstrate that the use of profiles improves homology detection and protein structure prediction. Profiles can be infered from protein multiple alignments using different approaches. The "Conservatism-of-Conservatism" is an effective profile analysis method to identify structural features between proteins having the same fold but no detectable sequence similarity. The information obtained from protein multiple alignments varies according to the amino acid classification employed to calculate the profile. In this work, we calculated entropy profiles from PSI-BLAST-derived multiple alignments and used different amino acid classifications summarizing almost 500 different attributes. These entropy profiles were converted into pseudocodes which were compared using the FASTA program with an *ad-hoc* matrix. We tested the performance of our method to identify relationships between proteins with similar fold using a non-redundant subset of sequences having less than 40% of identity. We then compared our results using Coverage Versus Error per query curves, to those obtained by methods like PSI-BLAST, COMPASS and HHSEARCH. Our method, named HIP (Homology Identification with Profiles) presented higher accuracy detecting relationships between proteins with the same fold. The use of different amino acid classifications reflecting a large number of amino acid attributes, improved the recognition of distantly related folds. We propose the use of pseudocodes representing profile information as a fast and

powerful tool for homology detection, fold assignment and analysis of evolutionary information enclosed in protein profiles.

INTRODUCTION

Homology detection is one of the central themes in bioinformatics because of its many applications in different areas, such as protein structure prediction (1). In many cases, proteins presenting the same fold have almost no amino acid sequence identity. This affects any protein structure prediction method based on homology detection; especially in the method's capacity to discriminate between true and false relationships.

Recently, different protein sequence comparison methods use profiles to improve detection of distantly related proteins. For instance PSI-BLAST (2) is an efficient and rapid method that detects homologous proteins by iterative building of position specific scoring matrices (PSSMs or profiles). However, when identity values are in the "twilight" or "midnight" zones, the relationships inferred from this analysis can be wrong (3). Besides, other technical complications exist: the number of iterations needed to find related proteins with the same fold within a database; selection of the initial scoring matrix to build a PSSM; or generation of a suitable PSSM using different protein seeds.

A profile that uses information contained in protein multiple alignments (4) is a powerful approach to analyze structural protein features. Mirny and Shakhnovich (5) prove that protein families presenting the same fold, but no evident sequence-based homology, have similar sequence entropy profiles. To perform their calculations, they classify amino

acids in six groups depending on their physicochemical properties. They argue that the distribution of important positions, either for the folding process or for protein stability, is conserved and can be detectable within the sequence entropy profiles derived from each sequence family (5)(6). This concept is known as the "Conservatism of Conservatism" (CoC).

To calculate a profile, it is important to choose a proper amino acid classification scheme. This is because the amino acid substitution frequency of a multiple alignment reflects the selection process and the evolutionary history of this particular protein family. However, there are many attributes to be considered when we classify amino acids. In a recent study that intends to solve the sequence metric problem (7), the authors perform a multivariate statistical analysis of a large number of amino acid attributes (almost 500) and summarize them in five multidimensional patterns. These five attribute co-variation factors reflect some properties with more weight, such as polarity, secondary structure, molecular volume, codon diversity, and electrostatic charge but many other properties are enclosed as well. The authors also suggest that it is easier to interpret and study evolutionary, structural, and functional aspects of protein variability using these factors.

In this work we used both the CoC entropy profile approach and the amino acid co-variation factors mentioned above, in a new method named Homology Identification with Profiles (HIP). Based on the comparison of entropy profiles derived from multiple alignments of homologous proteins, HIP identifies and groups protein families sharing the same fold. We calculated the entropy at each position (column) in multiple

alignments using the five multidimensional factors mentioned above, as well as physicochemical and structural amino acid classifications, to obtain a profile for each sequence family. We evaluated the entropy value frequency from each position in every protein family into twenty discrete ranges and translated them into pseudocode symbols based on the amino acid one-letter code. We used the FASTA program to perform multiple searches that detect similarity among entropy profiles from protein families of known structure. Finally, we selected the results with the best score from the searches performed in parallel using pseudocodes derived from each different amino acid schemes.

We compared HIP against other profile-profile comparison methods using a non-redundant subset of the CATH database (which stands for Class, Architecture, Topology and Homology)(8). The results showed that HIP reached coverage of ~30% of the expected results without errors. The use of the best result from any of the searches using the different amino acid classifications increased the accuracy of the method to detect related proteins sharing the same fold. This fact indicates that no single scheme can explain every family's evolutionary history.

HIP is a very accurate, profile-profile comparison method, which identified related proteins sharing the same fold. It also could be used for clustering sequence families with the same fold without needing any experimentally determined 3D information. We are currently investigating how the entropy profiles calculated with a wide scope of amino acid attributes can help us to understand the evolution of structure and function.

Methods

Databases and families of homologous proteins

We selected a subset of the CATH v2.6 database as the fold reference database (8)(9) (available at ftp://ftp.biochem.ucl.ac.uk/pub/cathdata/v2.6.0) filtering out sequences with over 40% identity, using the CD-HIT algorithm (10); domains with 20 amino acids or less were removed as well. Using each retained sequence as query, a PSI-BLAST (2) search was performed against a derivative UniProt Knowledgebase filtered at 90% sequence identity, (UNIREF90 database; available at ftp://ftp.expasy.org/databases/uniprot/uniref/) (11) containing 1,388,652 protein sequences. The search parameters were: BLOSUM62 scoring matrix; gap open/extension penalties of -11/-1; cutoff expectation value of 10^{-3} for inclusion in the next round and maximum of 5 iterations. We also stored the PSSM generated in the last cycle and extracted the final sequence alignment using Mview v1.47.3 (12). We built our homologous protein families using a maximum of 500 results from the final sequence alignments; all the alignments having less than 20 sequences were removed.

Amino acid classifications

To calculate the entropy values, we used twelve different amino acid classification schemes. We obtained ten of these amino acid classifications from Atchley *et al.* (7) where the sequence metric problem is addressed. They studied 494 amino acid attributes using a multivariate statistical analysis and summarized them into five different multidimensional scores or factors. For each of those five amino acid classifications, there was a dendogram (Factor I to Factor V) representation. From these dendograms, we obtained ten schemes for entropy calculation by using two different average distance cutoff lines (0.2 and 0.4) (Fig. 1). We also used the six classes of amino acids as reported by Mirny *et al.* (6) and the Lesk classification: small [AST], medium-sized and large hydrophobic [CVILPFYMW], polar [NQH], acidic charged [DE], basic charged [RK] and special [G]; we refer to these last classifications as MS and LESK respectively. Table I shows all twelve amino acid classifications.

Entropy calculation

We calculated the entropy value for each position in the multiple alignments as follows:

$$s(l) = -\sum_{i=1}^{n} p_i(l) \log p_i(l)$$

Where $p_i(l)$ represents the frequency of each of the i classes of residues at position l in the multiple alignment; n represents the number of amino acid groups depending the classification criteria.

Pseudocode calculation

Entropy values, calculated from each position of a multiple alignment, were translated into a pseudocode that uses the amino acid one-letter code. To do this, we binned the total entropy for each classification criteria in twenty intervals. We adopted the Equal Sized Bin (ESB) approach, where the size of the bin depends on the frequency of the entropy values, to make the intervals; all the values calculated, for all alignments from our CATH subset, were used to define the bin size for each classification. Each different pseudocode is named after the the binning condition (ESB) and the factor used for entropy calculation (Factor I-V, MS or LESK). The positions with 70% or more gaps were not considered for entropy calculation and were marked as gaps at the final pseudocode.

Pseudocode treatment

The twelve different pseudocodes, generated using the amino acid classifications, were analyzed for code complexity before performing comparisons among them. As we are working with letter codes, we evaluated their global complexity using the Shannon entropy informational theory (13) and removed all sequences having less than 2 bits. After this filtering process, 4,606 pseudocodes remained; for now on we will refer to this subset as the CATH40 database. Query pseudocodes were filtered before the searches with the SEG program (14) in order to remove pseudocode regions with low complexity.

HIPMAT matrix

We modified the idn_aa.mat identity scoring matrix, included in the FASTA (15) distribution, into a matrix called HIPMAT to evaluate the distance properties between the entropy intervals; we used values of 4 for interval identities on the matrix diagonal. For similarity values we reduced the value on the diagonal, proportionally to the double of the distance from one range to another in both directions. These values decayed from 4 to 2, then 1, 0 and then used negative values increasing by the same magnitude until reaching -4, which we kept as the highest penalty score for all of the most distant intervals. This system thus favors the scoring of similar entropy values among pseudocodes and as resemble FASTA matrices, it can be used to generate significant FASTA33 scores. Similar pseudocodes give rise to low E-values and conversely dissimilar pseudocodes produce high E-values.

HIP searches

We used the FASTA33 program (version 3.3t08d4, March 2001) (15). to compare the pseudocodes with the HIPMAT matrix. FASTA search parameters were set to default. We performed multiple searches using the twelve different pseudocodes described above. In each case we chose the best (lowest E-value) result, obtained from any of the twelve multiple searches, as the representative result for a given query.

Protein comparison methods

We used different sequence-profile or profile-profile comparison methods, to test the performance of our method against them. For each method, we used the same CATH40

database but with protein sequences. For sequence-profile comparison, we ran PSI-BLAST once using the previously defined PSSM for each query, with the same parameters described above. For profile-profile comparison, we used HHSEARCH and COMPASS algorithms to construct and compare profiles derived from PSI-BLAST multiple alignments, as described (1, 16). In the case of HHSEARCH, we added the secondary structure prediction (17) for each query when constructing its profile.

Coverage versus Errors per Query (CVE) plots

We performed Coverage versus Errors per Query (EPQ) plots (18) to compare the selectivity and sensitivity of HIP versus PSI-BLAST, COMPASS and HHSEARCH algorithms. For each method, its pair-wise comparison results are sorted from best to worst based on the score of interest (E-value). The coverage is increased if both query and target shared the same 4 CATH classification numbers (same homologous superfamily or homologues) otherwise the error is increased, except when the query and target shared the same 3 CATH numbers. Self-pairs results were discarded from the analysis for any of the methods. EPQ is the number of errors at a given value divided by the number of queries, and Coverage corresponds to the total of positive hits at the same value divided by the total number of possible homolog pairs.

Z-score calculation

We calculated a Z-score to evaluate if the fold distribution of the results and from our subset was significantly different. To calculate the Z-score we first calculated the

Standard Deviation (σ) between the observed distribution (results) and expected distribution (subset) using this formula:

$$\sigma = \sqrt{\Sigma (X_{obs} - X_{exp})^2 / n}$$

Where X_{obs} is the results fold distribution; X_{exp} is the subset fold distribution and n is the number of fold classes (4 in this case). For Z-score we used the formula:

$$\textbf{Z-score} = (X_{obs} - X_{exp}) / \sigma$$

RESULTS

CATH40 benchmarking results

To test the effectiveness of HIP as a remote homologue recognition system, we compared it to other sequence comparison methods commonly used to infer homology. We performed searches using HIP, PSI-BLAST, COMPASS and HHSEARCH by using the CATH40 database as the target set (see Methods). This library contains 4,606 entries representing 630 families of homologous superfamilies (homologues) with at least two members per family. Using the longest protein from each family as representative seed, searches were carried out using either protein sequences (PSI-BLAST, COMPASS and HHSEARCH) or pseudocodes (HIP) as queries.

To evaluate the results from the different comparison methods, we used CVE plots (see Methods). This representation allowed us to compare the results from different methods, independently of the reference value used by each one. HIP presented the best coverage

at a 0% EPQ level (Fig. 2). This is the most striking result since our method can relate proteins with the same fold without errors. PSI-BLAST, COMPASS and HHSEARCH, at the same EPQ level, had coverage of 3.67%, 5.77% and 3.67%, respectively. Table II shows coverages, for the rest of EPQ error levels and their corresponding E_values.

Contribution of amino acid classifications to selectivity

Our method presented a great selectivity in comparison to the tested methods. Selection of the best result for each query, from searches using twelve different factors, increased the method performance. We analyzed the distribution of true positive results for each factor (at a 0 EPQ level). Factor I classification found most of the true positive results, followed by MS and Factor V schemes (Fig. 3). When the average distance cutoff value was considered on each factor, the distribution wasn't affected. However, we observed that Factor I and V classifications with an average distance cutoff value of 0.4, presented a particular high frequency of true positive results (Fig. 4).

Detection of remote homologues at 0 EPQ rate

Table III presents the comparison of the results identity percentage ranges found by each method at 0 EPQ rate. HIP shows for every range the best performance and a considerable amount of results belong to low identity percentage rates (less than 20%). Interestingly, HIP can detect proteins with the same fold but low percentage identity with no errors in between. We analyzed the obtained results to evaluate the distribution of the type of fold classes and identity percentage rates in order to see if our method presented

any bias. Table IV presents the number of results obtained by HIP at 0 EPQ rate and their amino acid identity levels. The total number of results at 0 EPQ rate was 1,117, and 401 of them, (~35%) had 20% amino acid identity or less. Distribution of CATH fold classes (mainly alpha, mainly beta, alpha/beta and few secondary structures) at a certain range of amino acid identity percentage is also described.

To evaluate if our method presents a detection bias, we calculated a Z-score (see Methods) for each fold class. According to the Z-score values, no significant differences were found between the fold class distribution of results and the CATH40 subset (Table V).

We found 14 relationships that were not detected by other methods under the conditions and search parameters described. From these results, the relationship between the structures 1b73A1 and 1jflA2 corresponding to *Aquifex pyrophilus* glutamate racemase and *Pyrococcus horikoshii* aspartate racemase respectively is particularly interesting. These enzymes have around 13% identity. Previous reports (20, 21) suggest that these two enzymes have a common ancestor because of their conserved residues at the active site and the similarity of their structural elements. HIP search results, using both structures as queries against our CATH40 pseudocode database, are presented in Fig. 6. When the 1b73A1 pseudocode is used, the best match with the 1jflA2 pseudocode was found using the MS amino acid classification. On the other hand, the 1jflA2 pseudocode calculated with the Factor I (cutoff line 0.4), retrieves the best hit against the 1b73A1. It

is interesting that both pseudocode alignments report the same overlapping region with different amino acid classification as the best score.

Discussion

We presented here a new method for homology detection using protein sequence information obtained from multiple alignments. Our method takes the CoC analysis to a further level. In the CoC analysis, entropy values for each position are statistically evaluated, to find that the most conserved positions that are important for protein folding. Instead, we translate the entropy profiles into pseudocodes, to compare them as letter sequences using an existing program like FASTA.

For our purpose, we had to implement a scoring matrix which could measure the distance between the characters (representing entropy values) in the pseudocodes. The value of 4 on the matrix diagonal is a typical log-odd matrix calculation; this value is the result of the comparison of two identical codes. We didn't know *a priori* the frequency of changes between the letters of our pseudocodes, but we expected that the divergence pattern imprinted along the pseudocode will be conserved among families with the same fold. This means that, pseudocodes with the same fold would present the same pattern even if the entropy values differ in one or two entropy ranges. Another assumption is that very distant entropy values would have no relationship. With this reasoning, we built the HIPMAT matrix (see Methods) which allowed us to find that protein families with the same fold shared a common pattern of entropy values and not just the most conserved ones. Since we defined a code based on the same letters to represent amino acids, using

the FASTA program, thousands of searches can be performed in just a few minutes. The implementation of the HIPMAT matrix suitable for pseudocode evaluation allows us to use the FASTA program statistics.

In this study, we compare our method against other sequence-profile (PSI-BLAST) or profile-profile methods (COMPASS and HHSEARCH). HIP had better coverage than these methods at very low error rates, although all methods present a similar performance at error rates at 0.1 EPQ (~10% of errors). These results agreed with the known fact that sequence-profile and profile-profile comparison methods are efficient in the detection of homologous proteins, with the profile-profile methods showing a better performance in most cases (4). HIP reaches its coverage limit around the 0.1 EPQ level, this can be explained by the fact that a method exhibiting a great specificity looses sensitivity (19) but the coverage rate achieved with no errors is extremely good for an automated predictive method.

We analyzed the results obtained at a 0 EPQ rate in order to evaluate the number of pairs that, when compared using their amino acid sequence, presented identity values of remote homology (less than ~20%). As depicted in Table III, 401 results of the 1,117 would present 20% of identity or less, demonstrating that our method is capable of detecting a good number of remote homologues as well as relationships with higher identity percentages. This suggests that entropy profiles patterns can be used to detect similar folds without any structural information. The detailed statistical analysis and the correlation with the results with experimental folding kinetics data, indicate that positions

with strong CoC values are related to functional and structural features such as active site residues or key positions for protein folding (5).

One of the relationships detected only by our method is the pair glutamate racemace 1b73A1 and aspartate racemase 1jflA2. These domains share ~13% of amino acid identity in a Needleman-Wunsch global comparison of their amino acid sequences. We found this result notable because of the low identity between the protein domains and the phylogenetic distance between the source organisms (*Aquifex pyrophilus* [Bacteria] and *Pyrococcus horikoshii* [Archea]). Interestingly, previous reports propose that these to enzymes have a common ancestor and a similar evolutionary mechanism (21, 22, and 23). It has been suggested that the *P. horikoshii* OT3 aspartate racemase evolved through a complex evolutionary mechanism involving gene duplication, gene fusion and domain swapping (23). The pseudocode overlapping region of aspartate racemase (positions 62-91 and 206-237 of domain 1jflA2) over glutamate racemase (positions 106-138 of the 1b73A2 domain) is located in the middle of the two Rossman fold domains forming each protein. Our method found that only this overlapping region is necessary to relate these two proteins that present the same topology, despite their low sequence identity. We also expect that the evolutionary forces described for aspartate racemase could be applied to glutamate racemase. Currently, we are investigating more about these enzymes via the analysis of their entropy profiles, their metabolic role in each organism and the divergence in sequence and substrate recognition.

We propose that pseudocodes with the same fold, when their entropy profiles are aligned, not only positions with high conservation match but positions with similar divergence are also aligned, suggesting that the whole entropy pattern is conserved as a "fold signature".

The use of 12 different schemes of amino acid classification pseudocodes and the selection of the highest value obtained from all the searches, improved the specificity of the method. It is interesting that the detection of certain relationships were achieved only with certain amino acid classification. In most cases, the 12 searches reported a true positive value, but when the relationships became more distant, only some searches using pseudocodes with a particular amino acid classification, retrieved true positive results. These facts suggest that a family's evolutionary history cannot be explained using a single amino acid scheme and that different amino acids attributes have to be evaluated in order to detect certain relationships in spite of their sequence divergence.

In the CoC analysis, entropy profiles are calculated using a classification based on amino acid physicochemical properties. However, the use of this classification didn't retrieve most of the best score results. In fact, the scheme that found most of them was Factor I (Fig. 3); according to Atchley *et al*., this scoring factor has a significant correlation with a large set of different substitution matrices, suggesting a strong evolutionary basis associated to the multiple attributes related to this factor. In particular, the entropy calculation using the scheme described by Factor I with a cutoff value of 0.4 (Table I), presented a high frequency of the best score results (Fig. 4). None of the individual

searches using each amino acid classification scheme could show a better performance when compared with the HIP selection method (Fig. 5).

Additionally, we checked if our method had any bias in detecting a given fold class. We calculated a Z-score for the significance between the differences of the fold class distribution in our subset and our obtained results. A significant value of Z-score would be 3 or -3; all fold class Z-score results were well under these values spanning from -1.06 to 1.46, indicating that our method has no preference for any fold class. Likewise there was no fold class preference for non-detected families.

The analysis of which families are better recognized by particular amino acid scheme will help to understand the specific properties that describe protein relationships, and hence their evolutionary history. It is possible that the forces driving fold mechanisms or stability depend on a wide set of amino acid properties rather than on mostly structural ones. Further analysis of pseudocode-aligned regions is needed to achieve these goals. Also, an improvement of the distance matrix could be achieved by evaluating the frequencies of changes in a log-odd scoring calculation.

There are some other factors of high importance for the effectiveness of a profile-profile method in detecting distant homologues: the alignment quality; level of sequence divergence; gap content; and addition of structural or evolutionary information. When a multiple alignment contained very little information because of a reduced number of sequences or when the aligned sequences contained a lot of gaps or very little divergence

between the sequences, no informative pseudocodes could be generated. To avoid this limitation, we analyzed the informational content of our pseudocodes using the Shannon theoretical information theory (13) for a global analysis. To remove low complexity regions, we used the SEG program. These filters helped us to remove as many as possible false positive results. HIP cannot be used when a family of homologous proteins has few sequences (less than 20) or when there is a low degree of divergence between them, but we believe that with the growth of the protein sequence databases and enrichment with sequences from new species, these disadvantages will most likely disappear.

Any multiple alignments that fulfill the requirements to calculate a profile can be used as a pseudocode and be compared against a database of pseudocodes representing proteins with known structure. We have now the CATH40 database, a non-redundant, not overrepresented library of pseudocodes to perform searches against it, using any given query of unknown structure. We also have calculated pseudocodes for the complete CATH v2.6 database and for a UNIREF subset of sequences filtered to 40% of identity in order to perform further studies about protein clustering and fold assignment.

Conclusion

HIP recognizes relationships between proteins with the same fold with great accuracy. The results show that the use of pseudocodes, representing evolutionary information enclosed in protein sequence alignments, could be a fast and powerful tool to detect homology and fold assignment. Our method could easily be implemented as an automated protein structure prediction method, since it has a great trade-off between

sensitivity and selectivity. The simultaneous use of amino acid classifications that not only reflect structural properties but a large number of different attributes improves the recognition of distant related folds and should be further analyzed to study the evolutionary features present in proteins.

References

1. Söding J. Protein homology detection by HMM-HMM comparison. Bioinformatics 2005;21:951-960.

2. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997; 25:3389-3402.

3. Chung SY, Subbiah S. A structural explanation for the twilight zone of protein sequence homology. Structure 1996;4:1123-1127.

4. Ohlson T, Wallner B, Elofsson A. Profile-profile Methods Provide Improved Fold-Recognition: A study of Different Profile-profile Alignment Methods. Proteins 2004;57:188-97.

5. Mirny L, Shakhnovich E. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J Mol Biol 1999;291:177-196.

6. Mirny L, Abkevich VI, Shakhnovich E. How evolution makes proteins fold quickly. Proc Natl Acad Sci 1998;28:4976-4981.

7. Atchley,WR, Zhao J, Fernandes AD, Drüke T. Solving the protein sequence metric problem. Proc Natl Acad Sci 2005;102:6395-6400.

8. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH: A Hierarchic classification of protein domain structures. Structure 1997;5:1093-1108.

9, Pearl FMG, Lee D, Bray JE, Sillitoe I, Todd AE, Harrison AP, Thornton JM, Orengo CA. Assigning genomic sequences to CATH. Nucleic Acids Res 2000;28:277-282.

10. Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein database. Bioinformatics 2001;17:282-283.

11. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane N, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. UniProt: the Universal Protein Knowledgebase. Nucleic Acids Res 2004;32:D115-D119.

12. Brown NP, Leroy C, Sander C. Mview: a web compatible database search or multiple alignment viewer. Bioinformatics 1998;14:380-381.

13. Shannon CE. A mathematical theory of communication. Bell System Technical Journal 1948;27:379-423.

14. Wooton JC, Federhen S. Statistics of Local Complexity in Amino acid Sequences and Sequences Databases. Computers Chem 1993;17:149-163.

15. Pearson WR. Flexible similarity searching with the FASTA3 program package. Methods Mol Biol 2000;132:185-219.

16. Sadreyev R, Grishin N. COMPASS: A Tool for Comparison of Multiple Protein Alignment with Assessment of Statistical Significance. J Mol Biol 2003;326:317-336.

17. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195-202.

18. Brenner SE, Chothia C, Hubbard TJ. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci 1998; 95:6073-6078.

19. Sierk ML, Pearson WR. Sensitivity and selectivity in protein structure comparison. Prot Sci 2004;13:773-785.

20. Choi S, Esaki N, Makoto A, Yoshimura T, Soda K. Bacterial glutamate racemase has a high sequence similarity with myoglobins and forms an equimolar inactive complex with hemin. Proc Natl Acad Sci 1994;91:10144-10147.

21. Liu L, Iwata K, Kita A, Kawarabayasi Y, Yohda M, Miki K. Crystal structure of aspartate racemase from Pyrococcus horikoshii OT3 and its implications for molecular mechanism of PLP-independent racemization. J Mol Biol 2002;319:479-489.

22. Hwang KY, Cho C, Kim SS, Sung H, Yu YG, Cho Y. Structure and mechanism of glutamate racemase from Aquifex pyrophilus. Nat Struc Biol 1999;6:422-426.

23. Liu L, Iwata K, Masafumi Y, Miki K. Structural insight into gene duplication, gene fusion and domain swapping in the evolution of PLP-independent amino acid racemases. FEBS Letters 2002;528:114-118.

Acknowledgements

We thank Steven Brenner, Sarah Teichman, Kevin Karplus, Mark Dibley, Enrique Merino, and Alejandro Garcíarrubio for critical discussions and suggestions throughout the development of this work. William R. Atchley for providing us with amino acid classification data. David W. Krogmann for English and style corrections, Shirley Ainsworth for information and bibliography, and Arturo Ocadiz, Ricardo Ciria, and Jérôme Verleyen for help with computer systems. This work was supported by grants from CONACyT R33055N and DGAPA UNAM IN215201. ASF was supported by a CONACyT doctoral fellowship.

Table I. Amino acid distribution of the twelve factors used by HIP to calculate entropy profiles.

FACTOR/c #	AMINO ACID CLASSIFICATIONS											
MS	AVLIMC	WYHF	TQSN	RK	ED	GP						
LESK	AST	CVILWYMPF		HQN	RK	ED	G					
F-Ic2	AWM	GS	HPY	CVI	FL	DNQ	ER	K	T			
F-Ic4	AWM	GST	HPY	CVIFL	DNQ	ER	K					
F-IIc2	A	EM	L	CDT	NY	FIKH	QV	RW	G	S	P	
F-IIc4	A	EM	L	CDT	NY	FIKH	QV	RW	GS	P		
F-IIIc2	ACV	HPL	D	Q	S	ERGN	F	IMT	KW	Y		
F-IIIc4	ACV	HPL	DQ	S	ERGN	F	IMT	KW	Y			
F-IVc2	A	GT	LV	DKN	FQ	E	IPR	S	CMY	H	W	
F-IVc4	A	GT	LV	DKN	FQ	EIPRS	CMY	H	W			
F-Vc2	AWHC	G	LE	KY	IN	PV	MT	R	Q	D	S	F
F-Vc4	AWHC	G	LEPV	KYMT	INF	R	Q	D	S			

Abbreviations: MS = Mirny & Shakhnovich classification; LESK = Lesk structural classification; Factor/c # = Factor name / average distance cutoff value (See Methods).

Table II. Coverage reached by each method and its correspondent E-value

METHOD	*Coverage for EPQ rate / E-value*		
	0 EPQ	*0.01 EPQ*	*0.1 EPQ*
HIP	27.70% / 7.4E-13	28.74% / 4.1E-11	32.54% / 0.0011
COMPASS	5.77% / 2.55E-159	6.81% / 2.12E-150	33.31% / 4.14E-37
HHSEARCH	16.97% / 0	16.97% / 0	32.83% / 3E-24
PSI-BLAST	3.67% / 8E-56	17.30% / 3E-23	33.41% / 1E-07

Table III. Distribution of identity percentage ranges for results found by each method at 0 EPQ rate.

% ID Range (amino acid seq. comparison)	NUMBER OF RESULTS			
	HIP	COMPASS	HHSEARCH	PSIBLAST
$X \leq 10\%$	**32**	1	21	0
$10\% < X \leq 20\%$	**369**	58	254	11
$20\% < X \leq 30\%$	**535**	125	398	75
$30\% < X \leq 40\%$	**181**	50	137	55
Total	*1117*	*234*	*810*	*141*

Table IV. Distribution of folds and identity percentage ranges from results at 0 EPQ rate.

% ID Range (amino acid seq. comparison)	No. of results	Mainly Alpha	Mainly Beta	Alpha/Beta	Few Sec. Structures
$X \leq 10\%$	**32**	15	7	10	0
$10\% < X \leq 20\%$	**369**	91	90	185	3
$20\% < X \leq 30\%$	535	129	111	285	10
$30\% < X \leq 40\%$	181	40	30	97	14
Total (%)	*1117 (100%)*	*275 (24.62%)*	*238 (21.31%)*	*577 (51.65%)*	*27 (2.42%)*

Table V. Significance of the difference between database and results fold distributions.

Fold CATH Class	*CATH40 subset*	*Results 0 EPQ*	*Z-score*
Mainly Alpha	21.24%	24.62%	1.46
Mainly Beta	23.05%	21.31%	-0.78
Alpha/Beta	54.03%	51.65%	-1.06
Few Secondary Structures	1.58%	2.42%	0.37

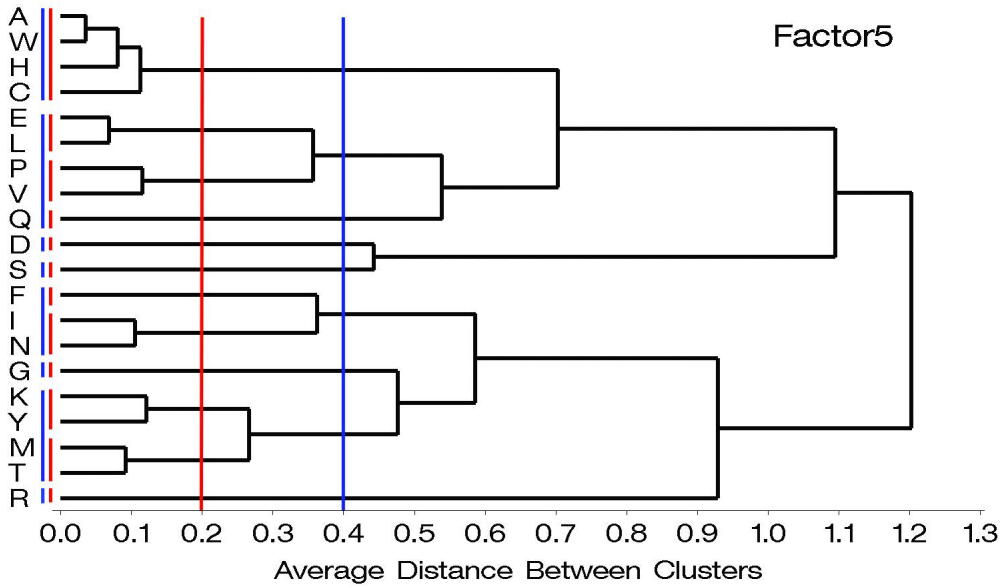


Figure 1: Example of classifications obtained from a factor dendogram. Amino acid clusters using 2 Average DistanceValue (ADV) cutoff lines. In red (dashed), clusters formed using a 0.2 ADV cutoff line; in blue (solid), clusters formed using a 0.4 ADV cutoff line.

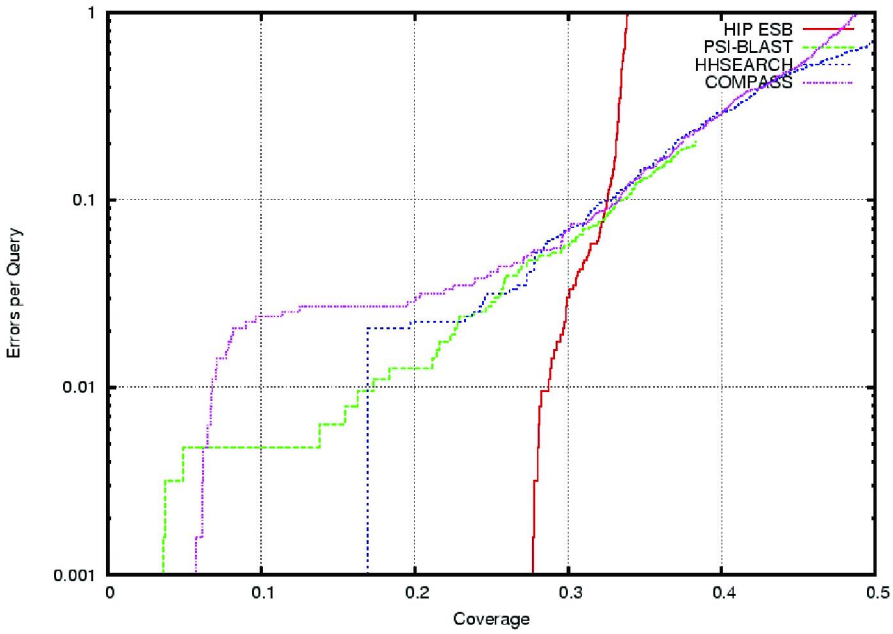


Figure 2: CVE plot comparison between HIP and other methods. An error per Query (EPQ) refers to the fraction of false positive (different CATH numbers) results of the total number of queries. Coverage represents the fraction of true positive (same 4 CAHT numbers) results of the total expected results (see Methods).

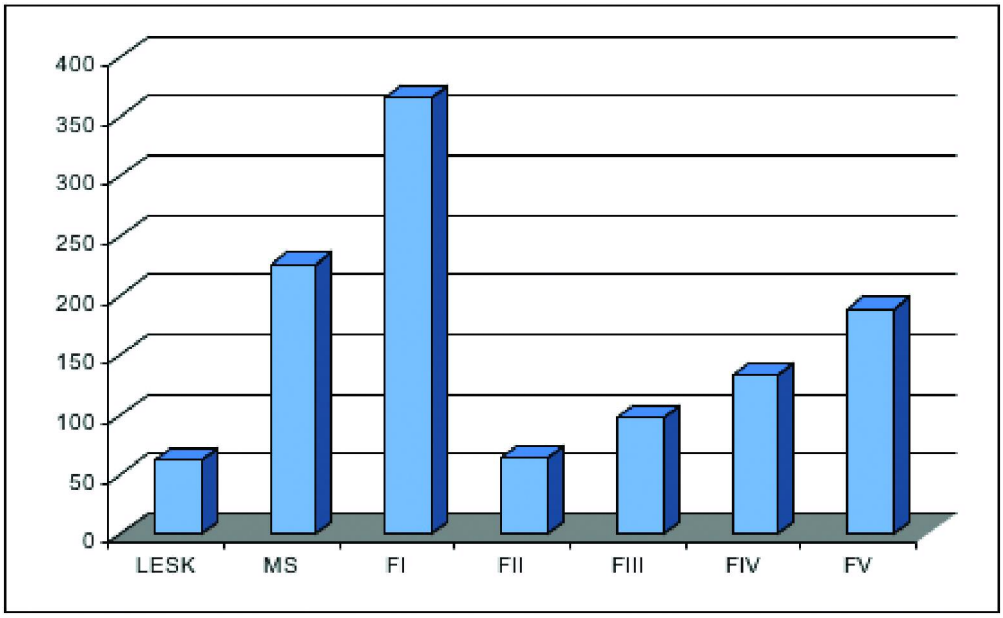


Figure 3: Distribution of results by classification factors. Y axis = Frequency of true positive results found at 0 EPQ level. X axis = Results grouped by factor.

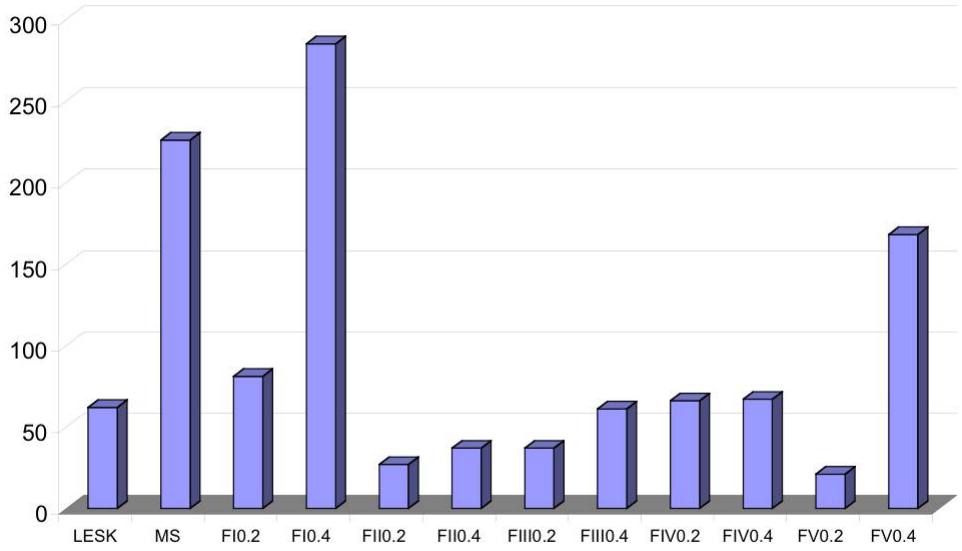


Figure 4: Distribution of results by factor and average distance cutoff value. Y axis = Frequency of true positive results found at 0 EPQ level. X axis = the twelve different amino acid classification.

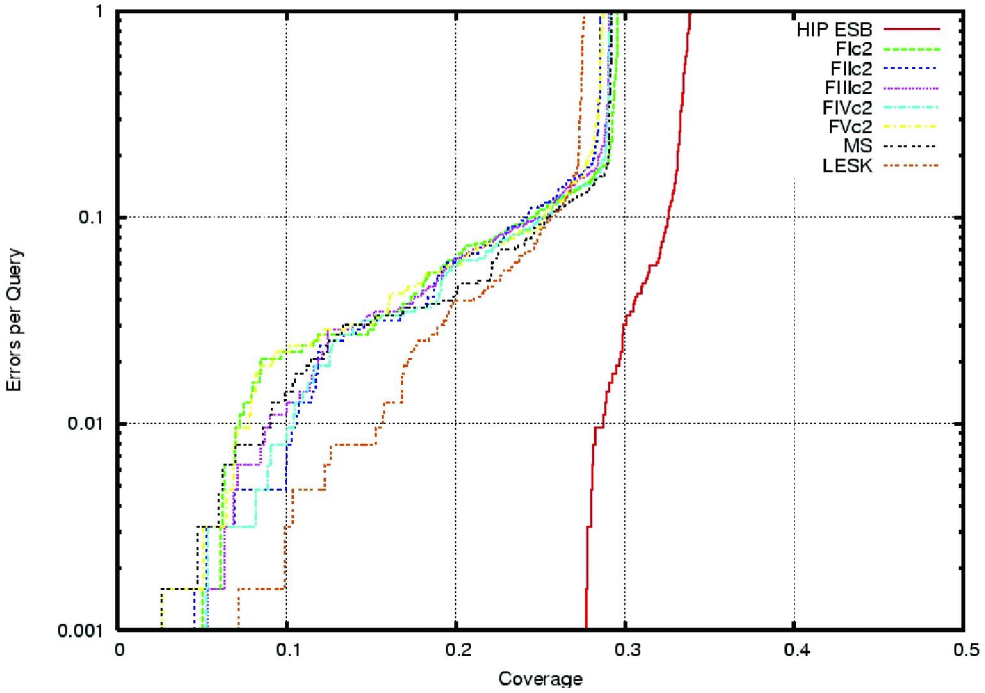


Figure 5A: CVE plot comparison between HIP and individual factor searches. Results from individual searches using the different amino acid classification schemes. Amino acid grouping using and average distance value of 2 as a cutoff line.

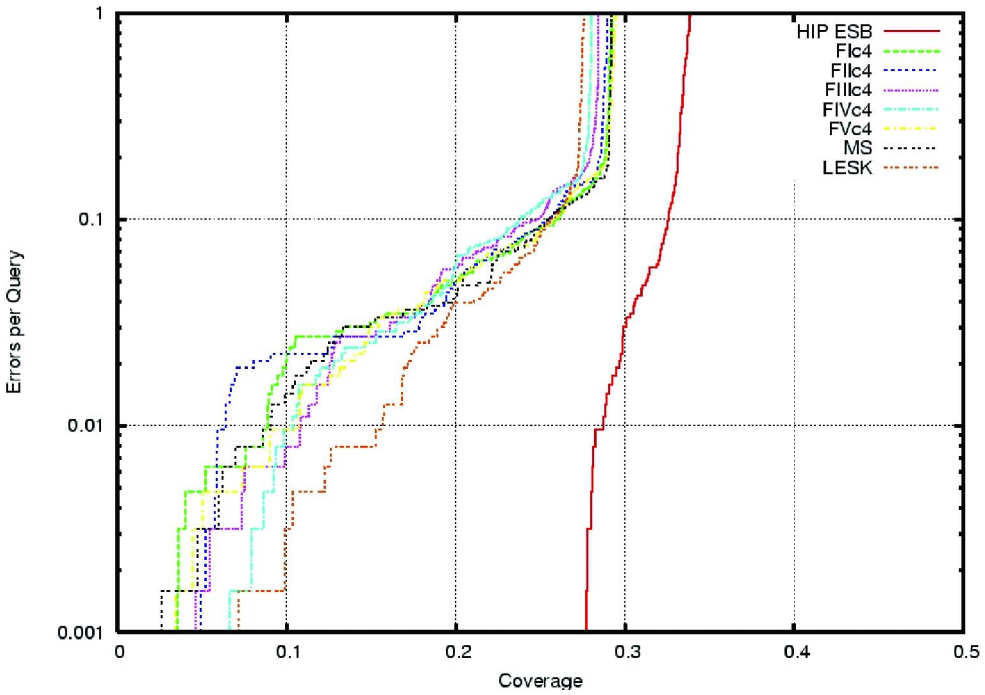


Figure 5B: CVE plot comparison between HIP and individual factor searches. Results from individual searches using the different amino acid classification schemes. Amino acid grouping using and average distance value of 4 as a cutoff line.

Search result of 1jflA2 vs CATH40 pseudocode database

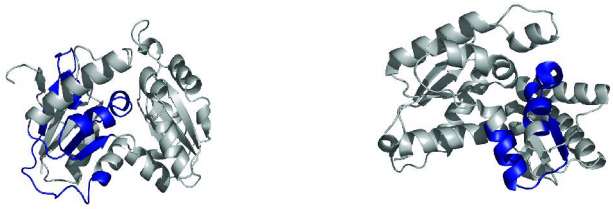
```
FASTA searches a protein or DNA sequence data bank version 3.3t08d4 Mar. 27, 2001
The best scores are:                                        opt bits E(5354)
1jflA2 479 3.40.50.1860                              ( 106)  245  155 1.5e-39
1b73A2 461 3.40.50.1860                              ( 106)  119   79 1.3e-16
1b73A1 455 3.40.50.1860                              ( 100)   94   64 4.2e-12


                                          10        20        30        40        50        60        70
1jflA2                                 TKHVVNSRRVDHADHAYQCHWPWSIYWVEYWSNQPLPKMSITQYWSMQTVPLIRGNWHSN**********
                                       .......:::::.:::::.:::::...
1b73A1 HFAAP PNTHSYKNTYMTPKCCLGKNPLVYLKYWPWSRVDHADGAYQCHWNWSIYWTFYYRSVTMVESVP
               40        50        60        70        80        90       100
```

Search result of 1b73A1 vs CATH40 pseudocode database

```
FASTA searches a protein or DNA sequence data bank version 3.3t08d4 Mar. 27, 2001
The best scores are:                                        opt bits E(5353)
1b73A1 455 3.40.50.1860                              ( 100)  208  130 3.9e-32
1b73A2 461 3.40.50.1860                              ( 106)   89   60 7.2e-11
1jflA2 479 3.40.50.1860                              ( 106)   66   46 9.2e-07


          70        80        90       100       110       120       130
1b73A1 CFCAAEGIKHGVYCKWWLWFPESLFIKMPPTGGTQITNQSCIADIAYPCDSMYPIYWNEYTPHNNRNDRLEHQ
                                    ....:.....::::.:.....:..::.....
1jflA2                           IEVKERQMSPRQDIADICYSDEVPYNGYWRFWWRRQQLNPEVRWPWWSIPVSIKTQFPYIYPYWPIWPHWYN
                                           10        20        30        40        50        60        70
```



1jflA 1b73A

Figure 6: HIP results for 1jflA2 and 1b73A1. The scores and alignments from the pseudocode FASTA searches are shown. The bidirectional results on each table are highlighted in bold. The overlapping region is colored in blue in both, the structure and alignment. Structure cartoons were drawn with PyMol (http://www.pymol.org)

BIBLIOGRAFÍA

1. Whitford,D. *PROTEINS. Structure and Function* (2005).

2. Patthy,L. *Protein Evolution*(Blackwell Publishing,1999).

3. Voet,D. & Voet,J.G. *BIOCHEMISTRY*2004).

4. Daggett,V. & Fersht,A.R. Is there a unifying mechanism for protein folding? *Trends in Biochemical Sciences* **28**, 18-25 (2003).

5. Kim,P.S. & Baldwin,R.L. Intermediates in the folding reactions of small proteins. *Annual Review of Biochemistry* **59**, 631-660 (1990).

6. Baldwin,R.L. How does protein folding get started? *Trends in Biochemical Sciences* **14**, 291-294 (1989).

7. Tompa,P. Intrinsically unstructured proteins. *Trends in Biochemical Sciences* **27**, 527-533 (2002).

8. Wood,T.C. & Pearson,W.R. Evolution of protein sequences and structures. *Journal of Molecular Biology* **291**, 977-995 (1999).

9. Mirny,L.A., Abkevich,V.I., & Shakhnovich,E.I. How evolution makes proteins fold quickly. *PNAS* **95**, 4976-4981 (1998).

10. Olmea,O., Rost,B., & Valencia,A. Effective use of sequence correlation and conservation in fold recognition. *Journal of Molecular Biology* **293**, 1221-1239 (1999).

11. Socolich,M. *et al.* Evolutionary information for specifying a protein fold. *Nature* **437**, 512-518 (2005).

12. Chothia,C. One thousand families for the molecular biologist. *Nature* **357**, 543-544 (1992).

13. Fitch,W.M. Homology: a personal view on some of the problems. *Trends in Genetics* **16**, 227-231 (2000).

14. Mount,W.D. *BIOINFORMATICS. Sequence and Genome Analysis*(CSHL Press,2001).

15. Henikoff,S. & Henikoff,J.G. Amino Acid Substitution Matrices from Protein Blocks. *PNAS* **89**, 10915-10919 (1992).

16. Dayhoff,M.O. Atlas of Protein Sequence and Structure. *National Biomedical Research Foundation* **5**, (1978).

17. Ohlson,T., Wallner,B., & Elofsson,A. Profile-profile methods provide improved fold-recognition: A study of different profile-profile alignment methods. *Proteins: Structure, Function, and Bioinformatics* **57**, 188-197 (2004).

18. Soding,J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951-960 (2005).

19. Chothia,C. & Lesk,A.M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823-826 (1986).

20. Jaroszewski,L., Rychlewski,L., Li,Z., Li,W., & Godzik,A. FFAS03: a server for profile-profile sequence alignments. *Nucl. Acids Res.* **33**, W284-W288 (2005).

21. Pearson,W.R. Flexible similarity searching with the FASTA3 program package. *Methods in Molecular Biology* **132**, 185-219 (2000).

22. Altschul,S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402 (1997).

23. Chung,S.Y. & Subbiah,S. A structural explanation for the twilight zone of protein sequence homology. *Structure* **4**, 1123-1127 (1996).

24. Mirny,L.A. & Shakhnovich,E.I. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *Journal of Molecular Biology* **291**, 177-196 (1999).

25. Atchley,W.R., Zhao,J., Fernandes,A.D., & Druke,T. Solving the protein sequence metric problem. *PNAS* **102**, 6395-6400 (2005).

26. Orengo,C.A. *et al.* CATH: A Hierarchic classification of protein domain structures. *Structure* **5**, 1093-1108 (1997).

27. Pearl,F.M.G. *et al.* Assigning genomic sequences to CATH. *Nucl. Acids Res.* **28**, 277-282 (2000).

28. Li,W., Jaroszewski,L., & Godzik,A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**, 282-283 (2001).

29. Apweiler,R. *et al.* UniProt: the Universal Protein knowledgebase. *Nucl. Acids Res.* **32**, D115-D119 (2004).

30. Brown,N.P., Leroy,C., & Sander,C. MView: a web-compatible database search or multiple alignment viewer. *Bioinformatics* **14**, 380-381 (1998).

31. Shannon,C.E. A mathematical theory of communication. *Bell System Technical Journal* **23**, 379-423 (1948).

32. Wooton,J.C. & Federhen,S. Statistics of Local Complexity in Amino acid Sequences and Sequences Databases. *Computers Chemistry* **17**, 163 (1993).

33. Sadreyev,R. & Grishin,N. COMPASS: A Tool for Comparison of Multiple Protein Alignments with Assessment of Statistical Significance. *Journal of Molecular Biology* **326**, 317-336 (2003).

34. Jones,D.T. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology* **292**, 195-202 (1999).

35. Brenner,S.E., Chothia,C., & Hubbard,T.J. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *PNAS* **95**, 6073-6078 (1998).

36. Sierk,M.L. & Pearson,W.R. Sensitivity and selectivity in protein structure comparison. *Protein Sci* **13**, 773-785 (2004).

37. Choi,S., Esaki,N., Ashiuchi,M., Yoshimura,T., & Soda,K. Bacterial Glutamate Racemase has High Sequence Similarity with Myoglobins and Forms an Equimolar Inactive Complex with Hemin. *PNAS* **91**, 10144-10147 (1994).

38. Liu,L. *et al.* Crystal Structure of Aspartate Racemase from Pyrococcus horikoshii OT3 and Its Implications for Molecular Mechanism of PLP-independent Racemization. *Journal of Molecular Biology* **319**, 479-489 (2002).

39. Hwang,K.Y. *et al.* Structure and mechanism of glutamate racemase from Aquifex pyrophilus. *Nat Struct Mol Biol* **6**, 422-426 (1999).

40. Liu,L., Iwata,K., Yohda,M., & Miki,K. Structural insight into gene duplication, gene fusion and domain swapping in the evolution of PLP-independent amino acid racemases. *FEBS Letters* **528**, 114-118 (2002).

41. Lo Conte,L., Brenner,S.E., Hubbard,T.J.P., Chothia,C., & Murzin,A.G. SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acids Res.* **30**, 264-267 (2002).