COMPARACIÓN TRIDIMENSIONAL DE LA CADENA POLIPEPTÍDICA DE PROTEÍNAS.

Aplicación de técnicas de visión computacional al análisis estructural de proteínas

CAPÍTULO V

V-A Introducción

El número de estructuras de proteínas en la naturaleza podría no ser ilimitado, como sugirió Chothia en 1992. Estimó que los organismos vivos podrían incluir no más de 1000-1500 familias de proteínas, a partir de las cuales divergerían todas las proteínas implicadas en la diversidad natural. En los últimos años se ha producido un importante incremento del número de secuencias y estructuras de proteínas en los bancos de datos correspondientes. Las herramientas, en forma de programas y algoritmos diseñadas para analizar estos bancos de datos, se han ido perfeccionando para realizar análisis exhaustivos y completos, especialmente de bancos de datos de secuencias. El análisis tridimensional de los bancos de datos estructurales, *Protein Data Bank* (PDB) con más de 9000 proteínas, están todavía lejos de poder ser llevados a cabo con total exhaustividad y garantía. El estudio de proteínas de estructura conocida se facilita enormenmente cuando existe relación secuencial entre ellas, siendo éste el caso del análisis de familias de proteínas. No obstante, cada vez más se describen en la literatura ejemplos de proteínas que muestran un andamiaje (*scaffold*) similar, incluso cuando sus secuencias no están relacionadas (Orengo et al., 1992 y 1994, Holm & Sander, 1994 y 1996, Chothia & Lesk, 1986; Pascarella & Argos, 1992). En estos casos comparar estructuralmente las proteínas resulta muy difícil, especialmente cuando tampoco existen estructuras secundarias regulares comunes.

El alineamiento de secuencias y de estructuras secundarias regulares son utilizados como elementos principales para realizar clasificaciones de proteínas en bases de datos como: SCOP, CATH o FSSP (Holm & Sander, 1996; Murzing et al., 1995; Orengo et al., 1997). En las dos últimas clasificaciones, las proteínas son agrupadas por métodos automáticos y se reflejan las tendencias estructurales de las familias de proteínas, pero no necesariamente las relaciones funcionales y evolutivas entre ellas, quedando ocultas a los métodos automáticos. Sin embargo, la clasificación SCOP se realiza en parte manualmente y tiene en cuenta los datos bibliográficos sobre las funciones de las proteínas. Este hecho permite encontrar esta clasificación dividida en tres grupos: proteínas homólogas cercanas, homólogas remotas y proteínas análogas (Murzing et al., 1995; Russell & Barton, 1994; Russell et al., 1997). El primer grupo incluye proteínas con similitudes secuenciales significativas entre ellas, mientras los dos últimos

grupos relacionan proteínas con poca o ninguna homología secuencial. Así, estas proteínas se dividieron en homólogas remotas (cuando tenían relación estructural y/o funcional y derivaban de un ancestro común), y proteínas análogas (cuando no estaban relacionadas por un gen ancestral).

El incremento del número de estructuras terciarias conocidas hace útil y necesario el desarrollo de métodos para la comparación de estructuras, paralelamente al desarrollo de los métodos de análisis secuencial. Esto se justifica porque los alineamientos estructurales de proteínas pueden descubrirnos propiedades estructurales y funcionales entre ellas que quedan ocultos a los alineamientos secuenciales. En la literatura se han descrito algunos casos, por ejemplo, la proteína codificada en el *Obese gene*, sobre la cual y tras su modelado se predijo un plegamiento tipo *helical cytokine* (Madej et al., 1995). Más recientemente, las relaciones estructurales y funcionales descritas entre el PCI y el EGF tras su superposición estructural podrían tener importantes consecuencias biomédicas (Blanco-Aparicio et al., 1998; Mas et al., 1998).

El principal problema para la comparación y superposición estructural entre dos proteínas es encontrar la correspondencia entre regiones de estas dos proteínas. En el modelado por homología se utiliza el alineamiento secuencial para establecer tales relaciones. Sin embargo, se ha descrito un importante número de casos donde un plegamiento (*fold*) presenta una secuencia claramente diferente (Holm & Sander, 1995). ¿Hay en el banco tridimensional de proteínas (PDB) motivos estructurales desconocidos? Realizar un análisis de todo el PDB es, hoy en día por los métodos disponibles, algo no asumible. Consecuentemente, algunas relaciones estructurales entre proteínas podrían quedar todavía ocultas. Los métodos más utilizados para la comparación de estructuras de proteínas emplean el calculo de RMSD (*Root Mean Square Deviation*), descrito anteriormente en el capítulo III, el cual puede fallar cuando la correspondencia estructural entre los residuos de las dos proteínas que se están comparando no es evidente. En los últimos años se han desarrollado varias metodologías para obtener comparaciones estructurales de proteínas no relacionadas. Sin embargo estas metodologías implican un gran número de cálculos y realizar un análisis global del PDB resulta imposible, además de estar sometidas a los errores propios del cálculo de RMSD. Otra importante limitación son los procedimientos matemáticos que se utilizan y las implicaciones

geométricas que suponen los algoritmos utilizados. Todos los métodos, como ya ha sido comentado, utilizan el cálculo de RMSD. Este cálculo implica dos operaciones geométricas:

a) la **traslación** de una proteína sobre la otra, generalmente producida por la superposición en el espacio de sus dos centros geométricos o de masas, y
b) la **rotación** de una proteína sobre la otra para minimizar la distancia media de los átomos equivalentes (véase el capítulo anterior para más detalle).

La rotación de las moléculas se realiza por un método iteractivo donde éstas se hacen rotar sobre sus centros de masas, previamente superpuestos en el proceso de traslación. Comúnmente se utilizan hasta 50 iteracciones que garantizan un buen resultado. No obstante, estas rotaciones dependen de la traslación. Cuando las proteínas que se están estudiando tienen una alta homología, los centros de masas o en otros casos los centros geométricos de ambas proteínas son puntos equivalentes[1]. En estos casos la traslación es correcta y por tanto la superposición de esas dos proteínas es también correcta. Cuando por el contrario las proteínas que se están comparando tienen baja homología, sus centros geométricos y/o de masas no son equivalentes, siendo incorrecta la traslación. Este efecto es obviado por algunos de los métodos que emplean el cálculo de RMSD ya que se van seleccionando trozos de proteínas para realizar la comparación estructural. A medida que los fragmentos comparados disminuyen de tamaño los centros de masas son más superponibles y por tanto se obtienen mejores superposiciones. La superposición de dos proteínas implica por tanto la fragmentación de ambas, y por tanto, el análisis total implica un cálculo global sobre los resultados obtenidos en el seccionamiento anterior. La Figura V.1 muestra un ejemplo donde la traslación desde los centros de masas es incorrecto.

[1] en el sentido de equivalencia de átomos utilizado durante todo este trabajo de tesis

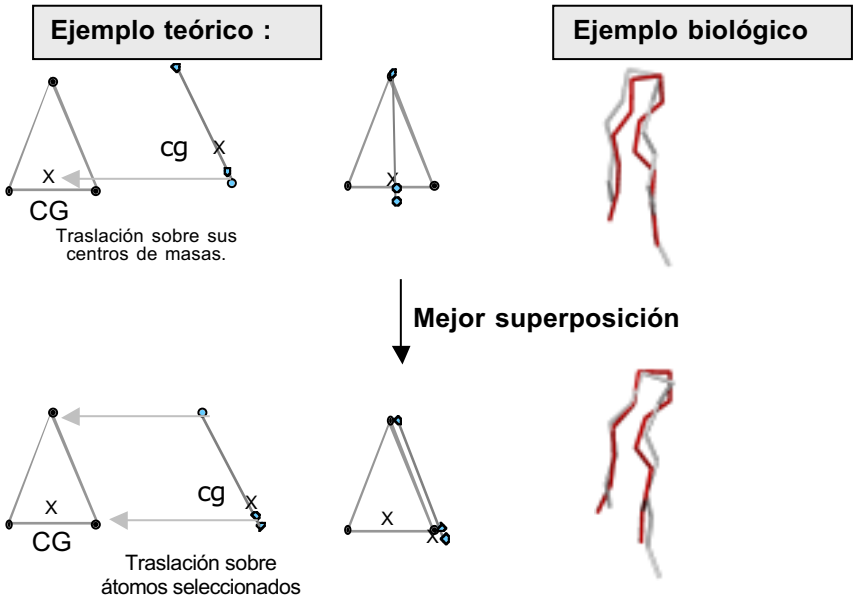


Figura V.1 Representación gráfica de dos posibles metodologías para la superposición de puntos.

V-B Objetivos

En este capítulo presentamos un método geométrico para la comparación estructural de proteínas. Este método permite obtener la correspondencia entre regiones de un modo **invariante a rotaciones y traslaciones**. Otra característica de la metodología que presentamos es la **flexibilidad** para elegir el detalle de análisis que el usuario desee. Así, el método permite comparar plegamientos generales (*folds*) de proteínas o elementos ultraestructurales como pueden ser lazos o elementos de estructura secundaria regular. Este efecto implica el uso de parámetros que permiten al usuario enfocar el grado de detalle de su interés, desde análisis burdos que comparan las proteínas enteras a comparaciones muy detalladas de estructuras o regiones pequeñas. Este efecto hace que el programa funcione de modo análogo a un microscopio, donde el usuario puede controlar la

lente o resolución, el campo de visión y el diafragma o sistema de contraste. Tal motivo hizo que llamáramos al programa *FOLD-SCOPE*.

Se han realizado estudios entre proteínas homólogas cercanas, homólogas remotas y análogas para contrastar el programa sobre el análisis de *folds* de proteínas. Se realizaron además análisis de ultraestructuras basadas en combinaciones de estructuras secundarias regulares y lazos, todos ellos contrastados con resultados publicados por otros autores.

FOLD-SCOPE es un programa realizado en Visual C++ preparado para ser ejecutado en máquinas IBM compatibles en entorno WINDOWS 9x o NT.

V-C Metodología

Para conseguir los objetivos planteados hemos desarrollado un método basado en técnicas utilizadas en el campo de la vision por computador. El método consiste en obtener las curvas que describen los plegamientos de las proteínas tomando tan sólo los carbonos alfa ($C\alpha$) y extraer de ellos una representación invariante a transformaciones rígidas (traslaciones y rotaciones). Obtenemos una representación que se mantiene constante independientemente de la posición y orientación en la que se encuentre, permitiendo de este modo la comparación directa de los *folds* de las proteínas. Se evita por tanto tener que definir la equivalencia entre los átomos de las dos proteínas, uno de los principales problemas que presentaban los cálculos de RMSD.

El método desarrollado permite regular la escala o grado de detalle aplicado en la comparación, adaptándose a comparaciones globales del *fold*, y a comparaciones detalladas en la que se tendrán en cuenta elementos ultraestructurales, como lazos o componentes de estructura secundaria regular.

En los tres puntos siguientes se describen las facetas en que se divide el método presentado.

Extracción de la representación invariante de los folds

Primero es necesario extraer la información que describe el esqueleto de cada proteína marcado por sus Cα. La secuencia de posiciones en el espacio de Cα de la proteína describe una curva que es ajustada a un β-*spline* cúbico. Se obtiene de este modo una curva tridimensional determinada por un conjunto de polinomios que describen el *fold* de la proteína. Este conjunto de polinómios se calcula a partir del conjunto de puntos discretos disponibles, los cuales son transformados extrayendo de ellos la magnitud de su **curvatura** y su **torsión**. Esta representación de la curva se mantiene invariante a transformaciones rígidas. De este modo, pasamos de una curva tridimensional (coordenada x,y,z), a una invariante y bidimensional (un valor de curvatura y uno de torsión) a la que llamaremos curva-TC.

Correlación de las representaciones invariantes

El objetivo de la correlación es comparar de forma eficiente las curvas-TC de las dos proteínas consideradas. Se identifican qué trozos de las dos curvas presentan similitudes estructurales, en qué grado, y se presenta a la vez un mecanismo de selección del nivel de detalle utilizado en la comparación. Básicamente la correlación se realiza muestreando las curvas-TC, es decir, pasando la curva expresada en forma polinómica a un conjunto discreto y ordenado de valores-TC. Los valores-TC son tomados de la curva a intervalos equiespaciados, determinados por el parámetro *SAMPLING* que expresa la distancia en angstroms del recorrido que hay entre muestras en la curva original.

El segundo parámetro que determinará el comportamiento o resolución del método es el *WINDOW-SIZE.* Este parámetro permite indicar cual es la longitud (número de muestras **N**) de la subsecuencia más pequeña que debe considerarse al hacer la comparación. Se calcula la correlación de la primera secuencia de **N** valores-TC de la primera proteína con la secuencia formada por todos los valores-TC de la segunda. Se tomarán los **N** siguientes valores-TC de la primera proteína y se repetirá el cálculo de la correlación para estos, y así sucesivamente. Este proceso se repetirá con todos las series de **N** valores consecutivos (ventanas) de la secuencia de la primera proteína.

El resultado del paso anterior, es una matriz de correlaciones **C**, en la que cada valor **N(i,j)** da una medida de la diferencia entre el trozo de la primera curva, de tamaño definido por el parámetro *WINDOW-SIZE,* y centrada en la posicion **i,** y el trozo del mismo tamaño de la segunda curva, centrado en la posición **j**.

Determinación de zonas geométricamente análogas a partir de los valores de correlación.

El tercer parámetro que regirá el funcionamiento de nuestro método es el umbral de aceptación (*threshold*), que indicará a partir de qué valor de correlación debemos considerar que dos subsecuencias (ventanas) presentan similitud estructural. Finalmente se asignan los resultados obtenidos a los Cα correspondientes a cada proteína.

Finalmente las secuencias de Cα candidatas son verificadas mediante el cálculo clásico de RMSD.

Se procede ahora a la descripción matemática de los algoritmos empleados en este programa.

Determinación de la función β-splines que describe el fold de la proteína

Una *curva* es una función vectorial de variable real, es decir, es una aplicación $r:\Re \to \Re^3$ que a cada valor real le hace corresponder un punto del espacio 3D (determinado por sus tres coordenadas **x,y,z**).

$$\dot{r}(s) = (x(s), y(s), z(s))$$

La curva que utilizaremos está parametrizada por la longitud del arco. La **longitud del arco** en un punto ***P*** de la curva es la distancia a la que se encuentra del inicio de la curva medida sobre ella. Es decir, la curva $r(s)$ en el punto *s,* nos indicara las coordenadas x,y,z de un punto medido sobre el arco a distancia s del inicio de la curva.

Determinamos las funciones **x(s), y(s)** y **z(s)** de modo que se ajusten a la colección de puntos ordenados $R = \{(x_1, y_1, z_1), (x_2, y_2, z_2), \mathrm{K}\ (x_N, y_N, z_N)\}$ que determinan nuestra proteína, es decir, a medida que incrementamos la variable ***s,*** los valores de nuestras funciones **x(s),y(s)** y **z(s)**, van recorriendo de forma suave los diferentes puntos **(x_i, y_i, z_i)** de la muestra ***R.***

Lo primero que hacemos es aproximar la longitud del arco entre dos puntos consecutivos (x_i, y_i, z_i) y $(x_{i+1}, y_{i+1}, z_{i+1})$ de la muestra calculando la distancia euclídea que los une.

$$\Delta s_i = \sqrt[2]{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2 + (z_{i+1} - z_i)^2}$$

De modo que ya conocemos ***N*** puntos de la curva $\dot{r}(s)$, ya que:

$$\overset{\mathrm{r}}{r}(\sum_{i=1}^{N} \Delta s_i) = (x_i, y_i, z_i)$$

Para describir el resto de la curva $\dot{r}(s)$ se han utilizado β-*splines* cúbicos. La aproximación por β-*splines* consiste en definir la curva por fragmentos, donde cada trozo está definido por un polinomio y determina la curva entre dos muestras consecutivas del conjunto de puntos. Dicho de otro modo, se definen los polinomios que unen los distintos Cα de la proteína y este conjunto de polinomios define el *fold* de la proteína.

El ajuste de la curva a β-*splines* consiste en determinar los coeficientes de los polinomios $P_i(s) = (p_{ix}(s), p_{iy}(s), p_{iz}(s)) \quad i \in [0, N-1]$ definidos entre cada par de puntos. En nuestro caso hemos utilizado polinomios de grado 3 (β-*splines* cúbicos). Para fijar los coeficientes de los polinomios se soluciona un sistema de ecuaciones en el que se impone: a) por ser β-*spline* cúbico, que el polinomio tiene que pasar por los Cα, y b) que en estos, la curva resultante tiene que ser C^2 (continua, y con primera y segunda derivadas continuas).

Cálculo de la torsion y curvatura.

Una vez que tenemos una forma analítica de la curva que pasa por los diferentes puntos de la muestra, procedemos a calcular su representación invariante a transformaciones rígidas, que será un descriptor diferencial de la misma (trihedro de Frenet).

Sea $\dot{T}(s)$ el vector tangente a la curva $\dot{r}(s)$ en el punto *s* (ver Figura V.2):

$$\vec{T}(s) = \frac{\partial \dot{r}(s)}{\partial s} = \left(\frac{\partial x(s)}{\partial s}, \frac{\partial y(s)}{\partial s}, \frac{\partial z(s)}{\partial s}\right)$$

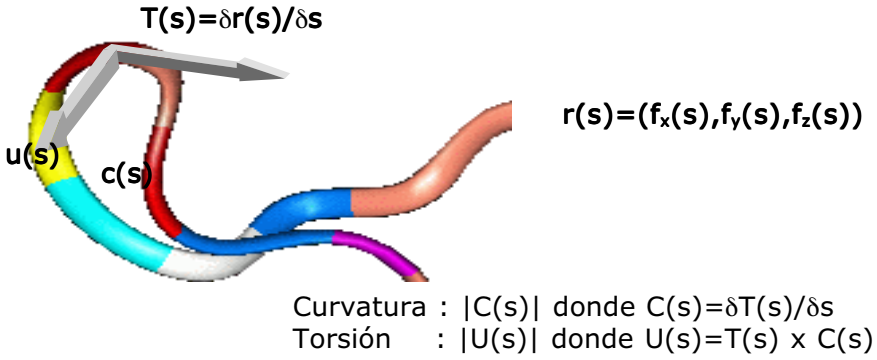


Figura V.2 Representación y descripción matemática de torsión y curvatura

El vector normal se describe como

$$\vec{C}(s) = \frac{\partial T(s)}{\partial s} = \left(\frac{\partial^2 x(s)}{\partial^2 s}, \frac{\partial^2 y(s)}{\partial^2 s} \frac{\partial^2 z(s)}{\partial^2 s}\right)$$

La función curvatura *k(s)* en cada punto de la curva $\dot{r}(s)$ será el módulo de esta función, por tanto será un escalar. Intuitivamente la función curvatura es inversa el radio del circulo que se inscribe en la curva en ese punto, así, si la curva es muy cerrada, el radio de la circunferencia que se inscribe es muy pequeño, y por tanto la curvatura es grande y viceversa. Se puede apreciar un esquema gráfico en la Figura V.3

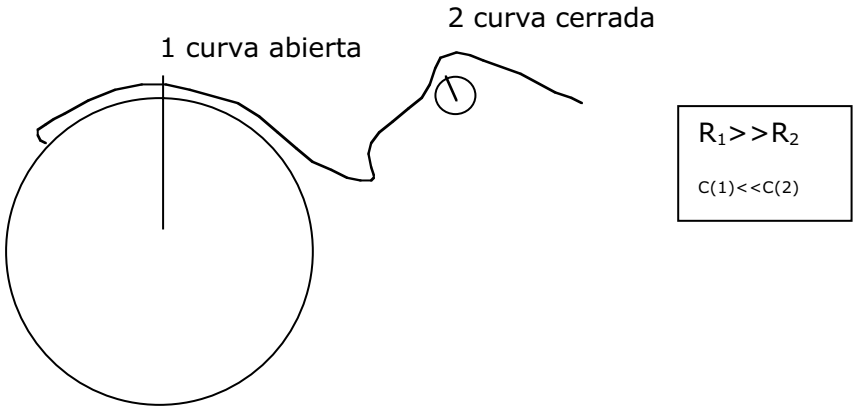


Figura V.3 Representación gráfica de la relación radio-curvatura

Matemáticamente,

$$k(s) = \left|\vec{C}(s)\right| = \left|\frac{\partial \vec{T}(s)}{\partial s}\right| = \sqrt{\left[\frac{\partial^2 x(s)}{\partial^2 s}\right]^2 + \left[\frac{\partial^2 y(s)}{\partial^2 s}\right]^2 + \left[\frac{\partial^2 z(s)}{\partial^2 s}\right]^2}$$

De modo que la tercera componente del trihedro de Frenet viene determinada por el vector torsión $\dot{U}(s)$. Éste es el producto vectorial del vector normal (curvatura) por el vector tangente. La magnitud de este vector *u(s)* puede verse intuitivamente como la variación de la curva respecto al plano determinado por estos dos vectores, es decir, si la curva tiene una tendencia a mantenerse o alejarse de dicho plano.

Su descripción matemática sería

$$\vec{U}(s) = \vec{T}(s) x \vec{C}(s) = \begin{vmatrix} \frac{\partial s}{\partial^2 x(s)} & \frac{\partial s}{\partial^2 y(s)} & \frac{\partial s}{\partial^2 z(s)} \\ \frac{\partial^2 s}{i} & \frac{\partial^2 s}{j} & \frac{\partial^2 s}{k} \end{vmatrix}$$

$$\vec{U}(s) = \left(\left(\frac{\partial y(s)}{\partial s}\frac{\partial^2 z(s)}{\partial^2 s}\right) - \left(\frac{\partial z(s)}{\partial s}\frac{\partial^2 y(s)}{\partial^2 s}\right), \left(\frac{\partial z(s)}{\partial s}\frac{\partial^2 x(s)}{\partial^2 s}\right) - \left(\frac{\partial x(s)}{\partial s}\frac{\partial^2 z(s)}{\partial^2 s}\right), \left(\frac{\partial x(s)}{\partial s}\frac{\partial^2 y(s)}{\partial^2 s}\right) - \left(\frac{\partial y(s)}{\partial s}\frac{\partial^2 x(s)}{\partial^2 s}\right)\right)$$

$$u(s) = \left|\dot{U}(s)\right|$$

Se puede ver que la curva $\vec{TC}(s) = (k(s), u(s))$ es invariante a transformaciones rígidas. Ésta es la curva transformada que utilizaremos para comparar las estructuras de proteínas. El hecho de que dispongamos de las curvas en forma polinómica permite realizar el cálculo de las derivadas analíticamente.

Muestreo de la curva TC

Han sido definidas la torsión y la curvatura como un conjunto de polinomios calculados a partir de las derivadas parciales de la curva original.

Para comparar las curvas se procede a muestrea las, es decir, extraer el conjunto de valores (valores-TC) $\{(k_0, u_0), (k_i, u_i), \mathrm{K}, (k_M, u_M)\}$ equiespaciados de la curva TC, es decir, de forma que la distancia entre muestras Δs (SAMPLING) sea la misma para todas las muestras, analíticamente:

$$k_i = k(i \cdot \Delta s)$$
$$u_i = u(i \cdot \Delta s)$$

De este modo, en lugar de tener una curva definida por una colección de polinomios, tenemos la curva definida por una colección de parejas de valores (k_i, u_i). Por la forma de construirlos sabemos además que la distancia entre los mismos es constante e independiente de la estructura proteica a partir de la cual hayamos determinado. Si comparamos dos *folds* provenientes de dos proteínas distintas pero con el mismo número de muestras, estaremos comparando dos curvas de exactamente la misma longitud de arco, ya que los puntos que la componen están exactamente a la misma distancia sobre la curva original.

Se puede ver como el parámetro Δs (SAMPLING) está relacionado con el grado de detalle y el tamaño de las comparaciones que queramos hacer.

Correlación entre secuencias de valores TC

La diferencia entre dos curvas en un punto *s* es la distancia euclídea entre los valores TC que le corresponden a la curva en ese punto:

Sea ***i*** la muestra correspondiente al punto ***s*** en la curva.

$$s = i \cdot \Delta s$$

sean $\{(k_0^1, u_0^1)(k_i^1, u_i^1)\text{K}, (k_M^1, u_M^1)\}$ y $\{(k_0^2, u_0^2)(k_i^2, u_i^2)\text{K}, (k_M^2, u_M^2)\}$ las secuencias de valores-TC de la primera y segunda proteína respectivamente.

Definimos la diferencia-TC en la muestra ***i*** como la distancia euclídea entre los dos valores-TC:

$$d_{i,i} = \sqrt[2]{(k_i^1 - k_i^2)^2 + (u_i^1 - u_i^2)^2}$$

Donde en general $d_{i,j}$ será la diferencia-TC entre la posición ***i*** de la primera curva y la posición ***j*** de la segunda.

Si en lugar de considerar la diferencia de las dos curvas en un punto, consideramos la diferencia entre las 2 secuencias de ***W*** muestras consecutivas definidas entre la muestra ***i*** y la muestra ***i+W*** de la primera proteína, y las muestra ***j*** y ***j+W*** de la segunda, calcularemos el valor medio de las diferencias-TC de las muestras consideradas, es decir:

$$d_{ij}^W = \frac{\sum_{k=1}^{W} d_{i+k,\, j+k}}{W}$$

Al número de muestras considerado en la comparación (**W**) lo llamaremos **tamaño de ventana** y a las muestras determinadas por este intervalo las llamaremos **muestras de ventana**.

El método de correlación considera cada una de las ventanas de valores TC de una proteína y calcula, de forma sistemática, la diferencia con todas las ventanas de la otra proteína.

Esto nos da como resultado una matriz de valores D donde cada valor D(i,j) será igual a la distancia d_{ij}^{W} .

Umbral de correlación y determinación de la secuencia coincidente más larga

Para determinar qué trozos son "geometricamente similares" se fija un umbral *threshold* o ***THD*** de aceptación. Así el fragmento de la primera proteína empieza en la muestra ***i*** de tamaño $W \cdot \Delta s$ y será considerada similar a la que empieza en la muestra ***j*** de la segunda si:

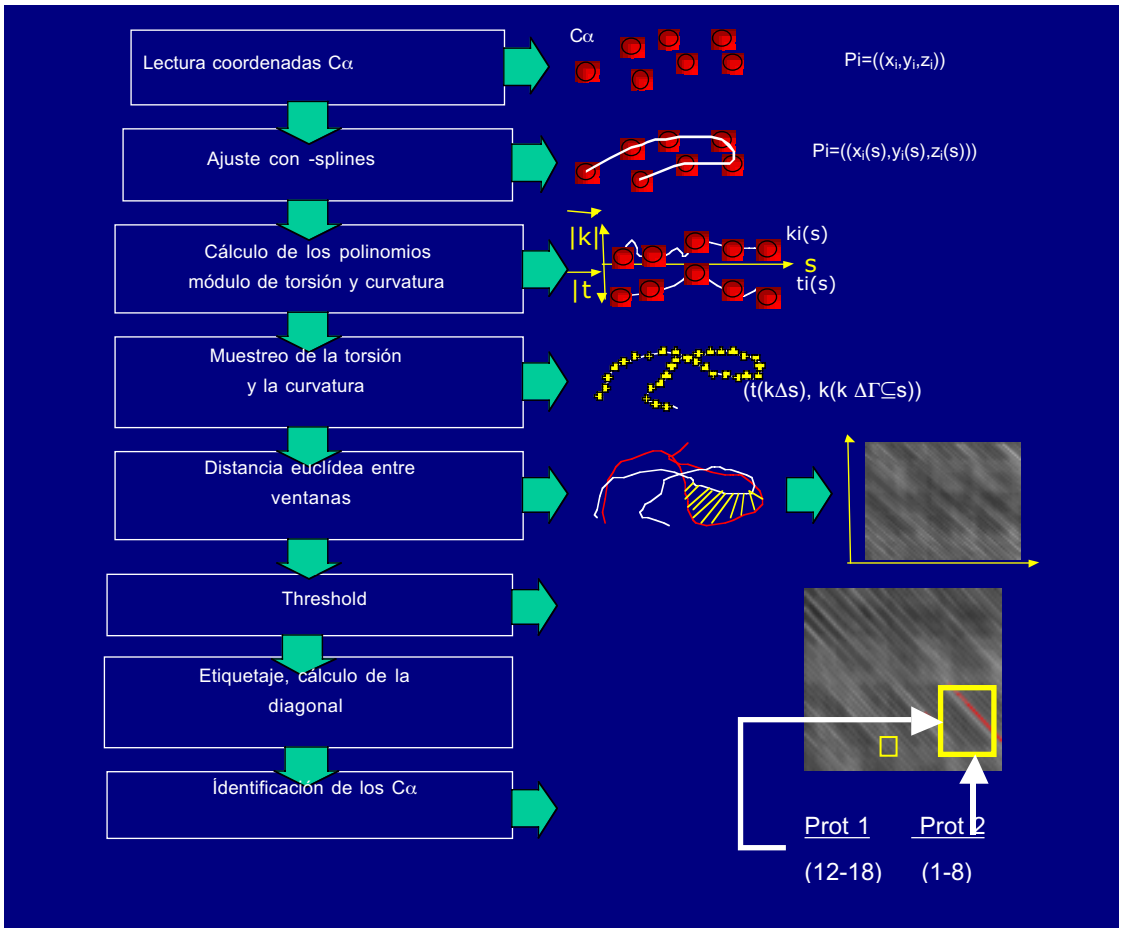
$$d_{ij}^{W} \geq THD$$

Una vez se tienen los puntos ***i,j*** en los que se ha considera que hay una secuencia similar, se busca la secuencia más larga en la que hay similitud, es decir, si se tiene que $d_{ij}^{W} \geq THD$ y al mismo tiempo $d_{i+1,j+1}^{W} \geq THD$, entonces la secuencia coincidente es la determinada por las muestras de la ventana que empieza en ***i***, más las muestra de la ventana que empieza en ***i+1***, es decir, que las muestras coincidentes serán (***i,i+1,i+2,...,i+1+W***) de la primera proteína y (***j,j+1,j+2,...,j+1+W***) de la segunda.

Para detectar las secuencias más largas se utiliza un método utilizado en visión por computador de extracción de diagonales máximas.

Finalmente cuando se tienen las secuencias de muestras coincidentes, se identifican los Cα que corresponden a cada proteína. Un esquema del funcionamiento del programa se puede ver en la Figura V.4

Figura V.4 Esquema del funcionamiento del programa



V-D Resultados

Para evaluar el programa FOLD-SCOPE se han llevado a cabo tres tests diferentes sobre la comparación de estructura de proteínas. Primero hemos chequeado la capacidad del programa para reconocer motivos estructurales formados por dos elementos de estructuras secundarias regulares unidos por un lazo. En segundo lugar, el programa ha sido utilizado para reconocer un dominio estructural incluido en el *fold* completo. Finalmente, el programa ha sido utilizado para comparar plegamiento completo de proteínas utilizando para ello

una proteína homológa cercana, una proteína homóloga lejana y una proteína análoga.

V-D.1 Comparación de motivos estructurales

Para la detección de motivos específicos hemos seleccionado del PDB proteínas que previamente han sido descritas en la literatura como contenedoras de motivos comunes (Oliva et al., 1997; Rufino et al., 1997; Wintjens et al., 1996 y 1998). Se han estudiado cuatro motivos diferentes (α/α, α/β, β/α y β-*hairpin*). Estos motivos han sido descritos por Oliva et al. en 1997; así se seleccionaron la clase 5.1.1 para los motivos α/α, el lazo de unión de Ca^{2+} (*EF-Hand)* para los motivos α/β, los P-*loops* de las proteínas que unen GTP para los motivos β/α, y finalmente, el lazo CDR-H3 de un anticuerpo para los *motivos β-hairpin*.

Motivos α/α

Las coordenadas del motivo α/α enmarcado en la clase 5.1.1. (Oliva et al., 1997) han sido extraídos de la proteína 1crl, entre los residuos 358 y 379. Han sido utilizados como elementos conformacionales para testear la capacidad del programa sobre las proteínas 1ede (*haloalkane dehalogenase*), 1rec (*recoverin*, una proteína con capacidad para unir calcio), 1s01 (*subtilisine*) y 7ccp (*cytocrome C peroxidase*). Estas proteínas han sido previamente relacionadas por Oliva et al. en 1997. El tamaño de ventana empleado es de 750 muestras, aproximadamente el tamaño total de la muestra, realizándose el muestreo cada 0,1 Å sobre un total de 80 Å. Los resultados de esta comparación se puede ver en la Tabla V.1 y en la Figura V.5.

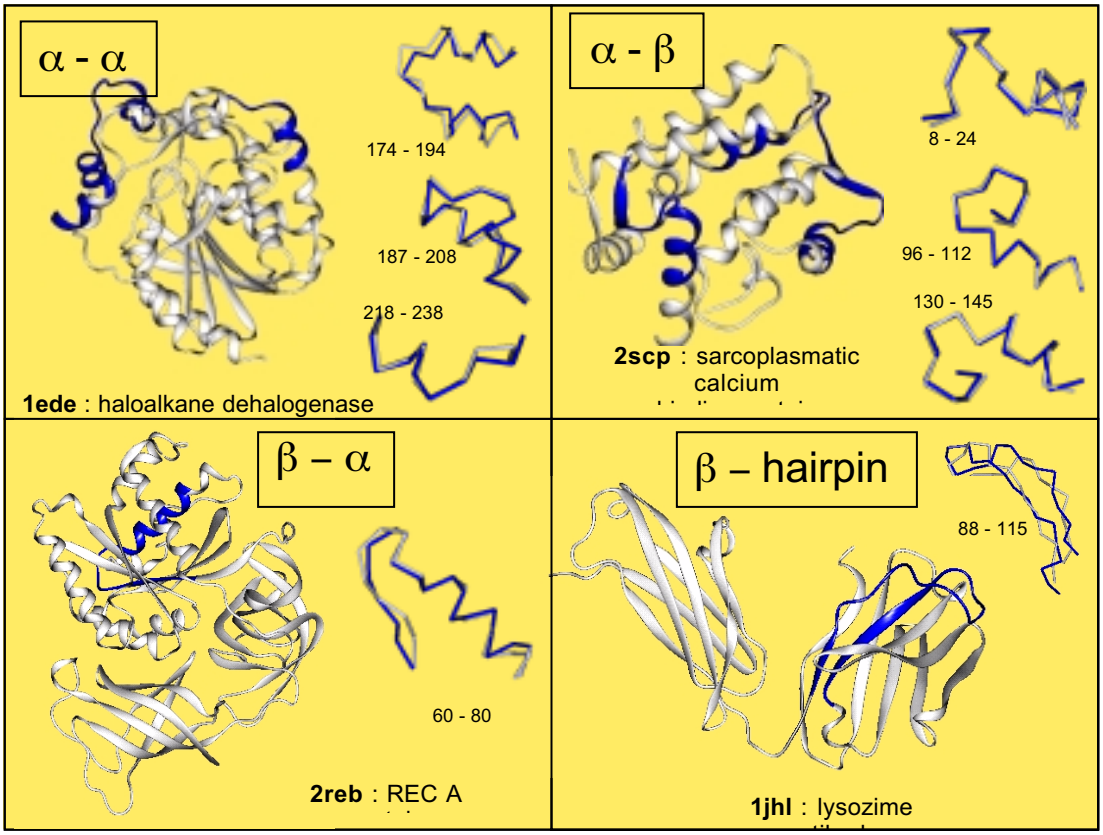


Figura V.5 Ejemplo de la comparación de motivos estructurales en proteínas.

Dos regiones con el mismo motivo estructural han sido detectadas para la proteína 1ede (residuos 174-194 y 187-208). Estas regiones presentan buenos valores de similitud (0,05 y 0,09), lo que equivale a un RMSD de aproximadamente 1,8 Å y 1,2 Å respectivamente. Las proteínas 1rec, 1s01 y 7ccp también presentan regiones con el mismo motivo estructural. Contrastando estos resultados con los previamente publicados en la clasificación de lazos (Oliva et al., 1997) podemos concluir que el programa es capaz no tan sólo de detectar los motivos estructurales para los que era testeado, sino que también reconoce motivos relacionados, como los residuos 146 al 168 de la proteína 7ccp que muestran un RMSD de 3,4 Å.

TARGET	PROTEINA (longitud)	MIN	ALINEAMIENTO	SS	RMSD (Å)	AUTOFOCUS (RMSD in Å)
α/α 1crl (80Å)	*1EDE (1171Å)*	*4.2*	*358-379 vs 174-194*	*0.088*	*1.84*	
			358-379 vs 187-208	*0.067*	*1.25*	
			358-379 vs 218-238	*0.111*	*6.37*	*367-377 vs 227-237 (1.1Å)*
	1REC	*23.3*	*358-379 vs 23-43*	*0.12*	*4.25*	
			358-379 vs 41-61	*0.101*	*2.71*	
	1S01	*4.6*	*358-379 vs 230-251*	*0.067*	*0.52*	
	7CCP	*10.4*	*358-380 vs 88-109*	*0.074*	*0.94*	
			358-379 vs 148-168	*0.118*	*3.38*	
			358-379 vs 243-263	*0.116*	*4.12*	
	Window Size 750Å, Sampling cada 0.1 Å, Threshold 30					
α/β 1rpt (61 Å)	*2sas (1171Å)*	*24*	*43-59 vs 62-77*	*0.111*	*1.64*	
			43-59 vs 122-137	*0.105*	*0.61*	
	2scp (1404Å)	*15.4*	*43-60 vs 8-24*	*0.195*	*0.51*	
			43-59 vs 96-112	*0.187*	*0.54*	
			43-59 vs 130-145	*0.202*	*0.67*	
	Window Size 575Å, Sampling cada 0.1 Å, Threshold 30					
β–α 1elf (97 Å)	*1gky*	*32*	*14-38 vs 4-24*	*0.137*	*1.29*	
	2reb	*21.6*	*14-38 vs 60-80*	*0.136*	*1.02*	
	5p21	*21*	*14-38 vs 6-26*	*0.139*	*1.20*	
	Window Size 730Å, Sampling cada 0.1 Å, Threshold 40					
β-hairpin 1vge (103 Å)	*8fab (799Å)*	*32.7*	*92-118 vs 91-117*	*0.139*	*1.34*	
	1jhl (433.8Å)	*34.5*	*92-119 vs 88-115*	*0.146*	*5.27*	*12-27 vs 99-114 (2.99Å)*
	Window Size 950Å, Sampling each 0.1 Å, Threshold 42					

Tabla V.1 La tabla muestra las proteínas que se comparan para cada motivo estructural. Los valores de MIN y SS hacen referencia a la similitud estructural de los elementos comparados (*ver texto*)

Motivos α/β

Los EF-hand han sido seleccionados para estudiar los motivos α/β. Las coordenadas de los motivos EF-hand pertenecen a las proteínas 1rpt (*prostatic*

acid phosphatase, E.C.3.1.3.2) entre los residuos 43 y 59, siendo utilizadas como conformación problema sobre la que se realizarán los distintos tests. Para chequear este tipo de ultraestructuras se utilizaron las proteínas 2sas y 2scp (*sarcoplasmatic calcium-binding proteins)* dado que en 1997 Oliva et al. habían descrito en estas proteínas tal motivo. El tamaño de la ventana utilizada es de 575 muestras tomando cada muestra a 0.1 Å del total de 60 Å que mide la conformación problema. Los resultados se muestran en la Tabla V.1 y en la Figura V.5. La proteína 2sas presenta el motivo EF-hand del mismo modo que la proteína 1rpt. La proteína 2scp presenta el motivo repetido tres veces (Tabla V.1). Estas regiones muestran buenos valores estructurales, muy por encima del resto de la estructura de la proteína. El RMSD calculado para estas regiones está entre 0,5 Å y 0,7 Å, lo cual confirma la capacidad del programa para identificar tales motivos estructurales.

Motivos β/α

Los P-*loops* han sido seleccionados para chequear este tipo de motivos estructurales. El motivo β/α de la proteína 1eft (*elongation factor TU)*, entre los residuos 14 y 34 ha sido utilizado como problema contra las estructuras 1gky (*guanuilate cyclase*), 2reb (*REC A protein*) y 5p21 (*oncogene protein*). El tamaño de ventana empleado fue de 730 muestras tomadas cada 0,1 Å. Las proteínas estudiadas mostraron unos valores de similitud estructural muy elevados, correspondientes a RMSD de 0,8 Å y 1,2 Å (ver Tabla V.1 y Figura V.5). Los resultados obtenidos fueron contrastados con resultados publicados en la literatura, confirmando el correcto funcionamiento del programa para caracterizar este tipo de motivos.

Motivos β*-hairpin*

Para chequear los motivos β-*hairpin* ha sido seleccionado el lazo del anticuerpo 1vge (TR1.9) frente a las estructuras de los anticuerpos 8fab (HIL) y 1jhl (D11.5). Se tomaron 950 muestras cada 0,1 Å. Los estudios previos de los *lazos* CDR-H3 muestran diferencias estructurales en el espacio de Ramachandran entre ciertos lazos de las proteínas 1vge y 1hjl. Sin embargo las comparaciones por metodologías específicas de lazos permitieron establecer relaciones estructurales entre ellos. Estas relaciones no han sido determinadas para la proteína 8fab

(Oliva et al., 1998). Los resultados obtenidos se muestran en la Tabla V.1. Estos resultados indican que aunque las conformaciones de Ramachandran sean distintas, la disposición de los Cα en el espacio muestran una conformación similar (Figura V.6). Consecuentemente hemos encontrado regiones equivalentes en los dos anticuerpos chequeados, lo cual muestra la capacidad del algoritmo incluso en casos de mayor dificultad, la comparación de los lazos CDR-H3 de los anticuerpos.

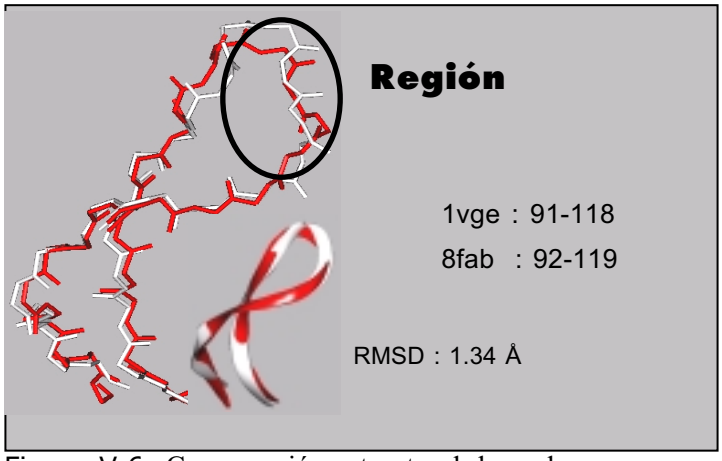


Figura V.6 Comparación estructural de un lazo

V-D.2 Reconocimiento de dominios estructurales

La siguiente cuestión planteada es: ¿puede el algoritmo/programa encontrar relaciones estructurales más complejas entre dos proteínas? Está claro que detectar el dominio de una proteína es más complejo que encontrar similitudes entre regiones enmarcadas entre dos estructuras secundarias regulares. Para chequear esta posibilidad del programa se ha utilizado el segmento de activación de la pro-carboxypeptidase B porcina (1pca B en código de PDB) frente a la pro-carboxypeptidase A . Para realizar el análisis se utilizó un tamaño de ventana de 100 muestras cogidas cada 0,75 Å sobre las 425 muestras posibles. Se obtuvo un valor MIN (mínimo de similitud) de 17 y se utilizó como *threshold* (o valor de corte) 25. La búsqueda permitió detectar dos regiones del segmento de activación comunes a ambas proteínas con valores de alineamiento estructural de 0,08 y 0,09 cuyos RMSD correspondían a 1,7 y 2,9 Å respectivamente. Estos resultados pueden verse en la Figura V.7.

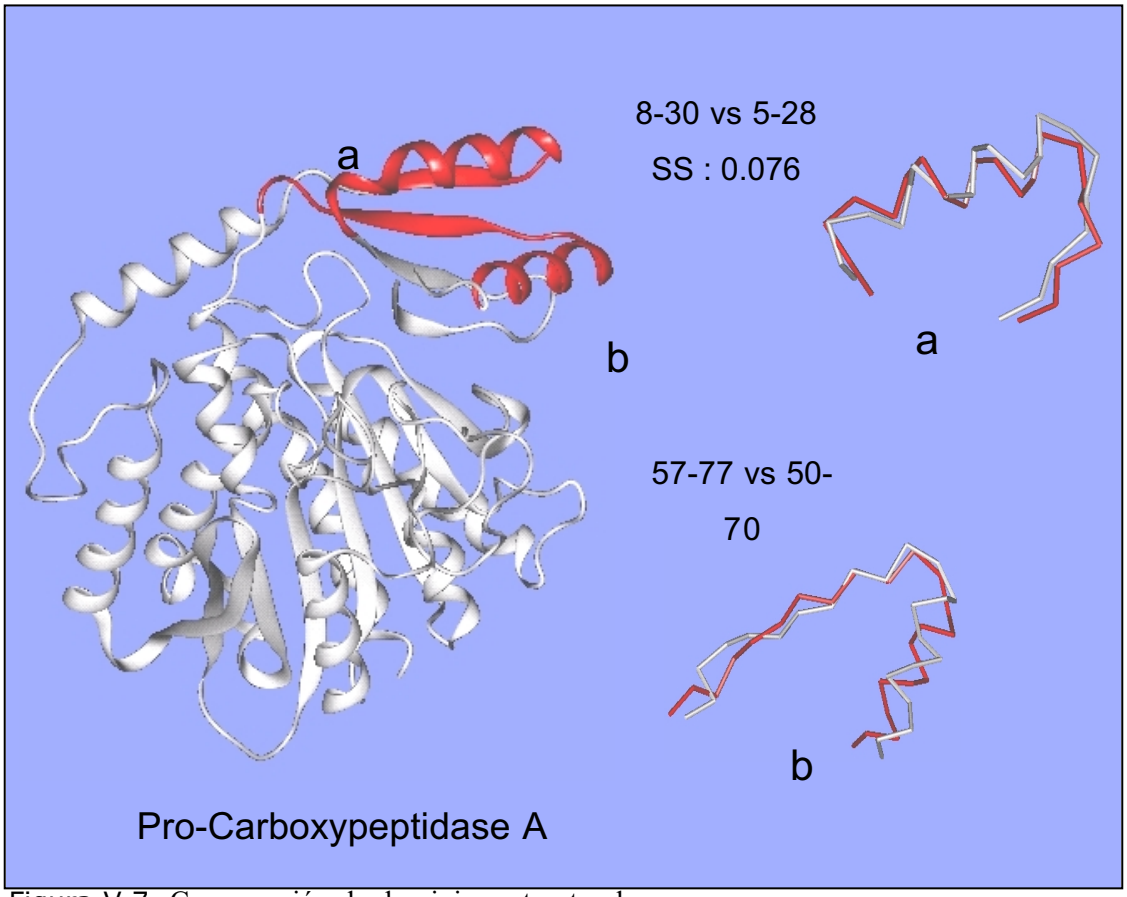


Figura V.7 Comparación de dominios estructurales

Los resultados muestran la capacidad del programa para detectar formas topológicas complejas, como son los dominios, en el conjunto de una proteína. Para ello tan solo cabe ajustar las condiciones de análisis del programa.

V-D.3 Reconocimiento completo entre proteínas relacionadas

Como consecuencia del apartado anterior nos planteamos la posibilidad de si el programa sería capaz de detectar diferencias estructurales de *folds* de proteínas similares. ¿Podría el programa distinguir entre proteínas homólogas cercanas, proteínas homólogas remotas y proteínas análogas?

La estructura del *OB fold* que se encuentra en la enterotoxina 1lts-d ha sido seleccionada para realizar este estudio. Nuestros resultados serán comparados

con los resultados obtenidos por Russell et al en 1997 que utilizaron tal estructura para realizar un estudio similar basado en datos secuenciales. La Figura V.8 muestra el *fold* de la enterotoxina junto con una proteína homóloga cercana, (1chp-d, *cholera toxin*) con un 80% de homología secuencial, una homóloga remota (1tss-a1, *chock syndrome toxin*) con un 8,8% de homología secuencial y finalmente un proteína análoga (1krs, *tRNA syntetase*) con tan solo el 4,4% de homología secuencial.

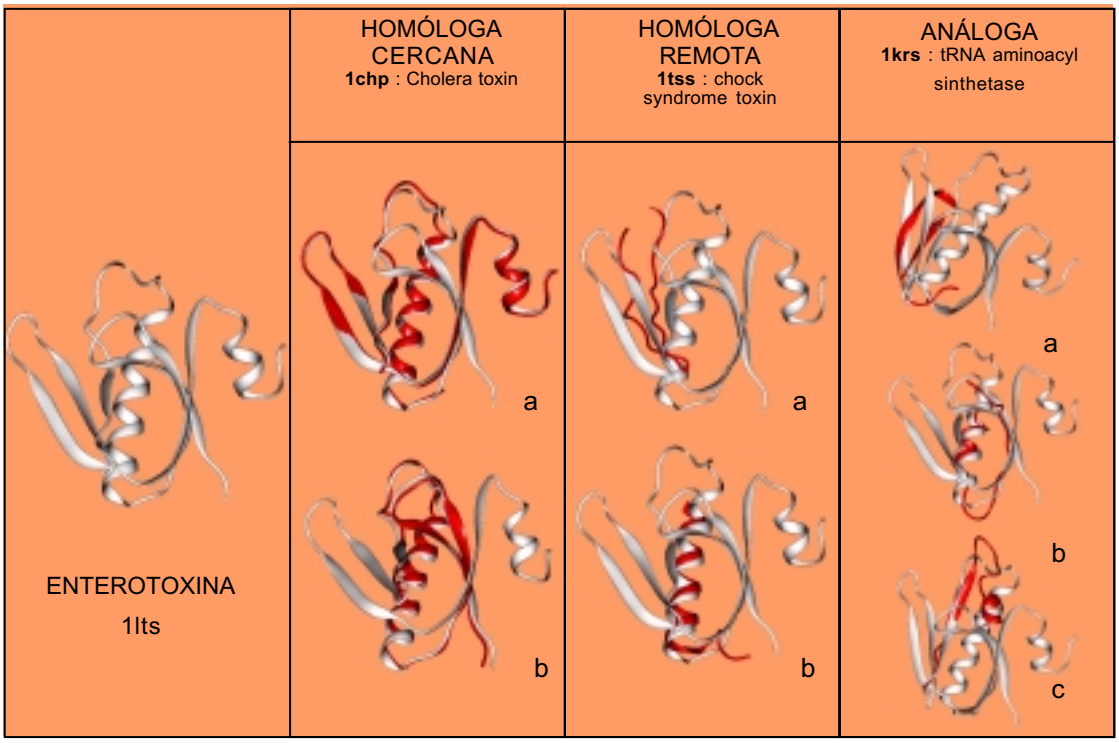


Figura V.8 Comparación estructural de una proteína con una homóloga cercana, una homóloga remota y una proteína análoga.

La estructura de la enterotoxina se utiliza como proteína problema frente a las estructuras de las otras tres proteínas. El muestreo se realiza cada 0,75 Å sobre una longitud total de 388 Å. El tamaño de la ventana utilizada es de 200 Å, aproximadamente el 70% del total del arco descrito por la proteína 1lts-d. Los resultados mostraron un valor mínimo (MIN) de 11,3 para la homóloga cercana, 58 para la homóloga remota y 91,5 para la proteína análoga (ver Tabla V.2). El *threshold* empleado para el análisis se situó entre el 75% y el 85% del valor MIN. Este resultado indica que es posible reconocer la relación estructural entre proteínas a partir de los valores obtenidos sin necesidad de realizar ningún cálculo de RMSD, es decir, a de los valores MIN. Además, las regiones con

mayor equivalencia estructural, más similares, quedaban resaltadas por el autofocus del programa (Tabla V.2). Tras el cálculo de RMSD, aquellas regiones que mostraban valores de RMSD superiores a 5 Å fueron automáticamente seccionadas por el programa. De este modo el programa actuaba como una máquina capaz de auto-enfocarse, mostrando qué partes de estas regiones tenían mayor similitud estructural. Los resultados muestran valores de RMSD de 1,1 Å y 4,5 Å (Tabla V.2) para el conjunto de las comparaciones. De este modo las proteínas homólogas fueron relacionadas en casi un 80% de su longitud total mientras la proteína análoga era relacionada sobre el 25% de su longitud, Tabla V.2.

FOLDS SIMILARES							
TARGET (387.7Å)	PROTEÍNA (longitud)	MIN	T	ALINEAMIENTO	SS	RMSD	AUTOFOCUS (RMSD)
1lts (387.7Å)	1chp (387.3Å) CERCANA	11.3	15	1-81 vs 1-81	0.065	1.1	
				44-103 vs 44-103	0.067	1.2	
	1tss (733.9Å) REMOTA	58	68.5	18-58 vs 51-91	0.27	11.05	33-54 vs 66-87 (3.05Å)
				33-84 vs 96-147	0.25	8.03	62-83 vs 125-146 (2.5Å)
	1krs (408.2Å) ANÁLOGA	91.5	109.3	19-73 vs 27-82	0.39	6.7	21-41 vs 30-50 (4.6Å)
							45-68 vs 53-77 (4.1Å)
				62-103 vs 67-108	0.42	6.8	65-89 vs 67-41 (3.14Å)
Tamaño de ventana empleado 200Å, muestreo cada 0.75Å							

Tabla V.2

Finalmente se ha aplicado un análisis de dos proteínas no relacionadas estructuralmente, el PCI (*Potate Carboxypeptidase Inhibitor*) y el EGF (*Epidermal Growth Factor*). Estas proteínas habían sido relacionadas estructuralmente por la topología de sus puentes de azufre y contrastados experimentalmente (Blanco-Aparicio et al., 1998; Mas et al., 1998), tal y como se muestra en el capítulo II. Sin embargo, los análisis realizados hasta ahora que abarcan la arquitectura global de estas proteínas y las relaciones estructurales/funcionales se habían realizado de un modo manual, sujetos por tanto a no ser completos. El uso del FOLD-SCOPE nos ha permitido relacionar más claramente estas proteínas,

estableciendo mayores similitudes estructurales entre ambas. Los resultados obtenidos se presentan en el último apartado de esta tesis junto con un estudio más detallado de ambas proteínas.

V-E Discusión y conclusiones

Proponemos en este capítulo de tesis un nuevo método (un algoritmo y un programa FOLD-SCOPE) para comparar estructuras de proteínas. Este método funciona de un modo análogo a un microscopio, detectando regiones comunes entre proteínas en función del grado de resolución que se aplique al análisis (desde elementos ultraestructurales como lazos hasta plegamientos enteros de proteínas). Los resultados presentados demuestran que el programa permite detectar relaciones estructurales entre proteínas y establecer relaciones evolutivas entre ellas en los casos ensayados. En el caso de corroborar este efecto este programa facilitaría una parte del trabajo a los actuales métodos de clasificación de proteínas.

Este trabajo demuestra que el valor de RMSD comúnmente utilizado es superado por el algoritmo que proponemos para comparar estructuras no relacionadas. Por otra banda, el cálculo del valor de RMSD entre dos estructuras es muy costoso y limita su uso en la comparación global del PDB. Para realizar un cálculo de RMSD es necesario definir la equivalencia de los átomos que se están comparando, siendo éste un ejercicio complejo y fuente de uno de los principales problemas a la hora de comparar dos estructuras de proteínas.

Los métodos comunes para la superposición estructural de proteínas obtienen la correspondencia entre dos *folds* a partir de la información secuencial, o a partir de matrices de distancias (Taylor & Orengo, 1989). La mayoría de estas aproximaciones fallan cuando es necesaria la aceptación de *gaps* (regiones sin correspondencia). Un efecto directo de este problema se observa en el *inverse protein folding* (reconocimiento del *fold* a partir de la secuencia) o en el *fold recognition* (reconocimiento del tipo de *fold*). Nuestra metodología permite mostrar la correspondencia estructural de dos *folds* en general y posteriormente

establecer cuáles son las regiones que presentan una equivalencia estructural mayor, y por tanto, que contienen los valores menores de RMSD.

Esta metodología puede ser utilizada para detectar los elementos claves de interacción de las proteínas con receptores o con otras proteínas, localizando en las zonas de interacción aquéllas que muestran la mayor diferencia estructural y que por tanto son las zonas que definen las interacciones. El programa puede ser utilizado para detectar posibles funcionalidades de regiones de proteínas. Los análisis secuenciales empleados para tal fin pueden ser complementarios y no son capaces de establecer ninguna relación entre dominios cuando la homología sencuencial es baja. Un claro ejemplo de lo anteriormente expuesto se muestra en el último capítulo de esta tesis por lo que se refiere a la comparación del EGF y del PCI. A partir de la detección de regiones con equivalencia estructural entre proteínas se puede inferir sobre posibles funciones heterólogas de las mismas. En este sentido el análisis de estructuras de péptidos unidos a anticuerpos marcan la estructura que reconoce tal receptor. Obviando la no poca importancia de las cadenas laterales, es posible detectar regiones en proteínas que presenten la misma estructura que muestra ese péptido, es decir, permite la detección de posibles epítopos, desde un punto de vista estructural, de una proteína. Un sencillo análisis secuencial complementario de los epítopos marcados podría ser suficiente para tal aplicación del programa.

La capacidad de autoenfoque del programa permite además el análisis completo de la proteína, detectándose fácilmente las regiones con similitudes estructurales comunes a las dos proteínas. Además el programa puede ser utilizado de un modo análogo a un microscopio por el usuario, permitiéndole enfocar elementos muy pequeños o rastrear la proteína de un modo más burdo.

Implicaciones biológicas

Las implicaciones biológicas que se pueden extraer de este estudio son una por cada grado de profundidad del análisis estructural. El caso del lazo de los CDR-H3 es uno de los más interesantes ya que muestran la capacidad analítica del programa sobre ultraestructuras de proteínas. La comparación de proteínas no relacionadas, como el EGF y el PCI, cuyos resultados se presentan en el capítulo siguiente, es el otro punto de gran interés al tratarse del análisis comparativo de dos *folds* completos. Finalmente, el diseño de una estrategia ágil y exhaustiva para comparar proteínas ofrece un amplio abanico de posibilidades en el campo de la biocomputación, ya que puede ayudar en la clasificación y relación entre proteínas con baja homología secuencial.

El trabajo realizado sobre el motivo CDR-H3 permite obtener una clasificación de este motivo, pero además permite establecer posibles patrones de reconocimiento, que apoyado con sencillos estudios secuenciales, abren otra posible vía de análisis. Un claro ejemplo de esto sucede al comparar las proteínas 1vge, 1jhl y 8fab. Los resultados indican que las conformaciones de los lazos de estas tres proteínas eran solapables, mientras el completo análisis realizado por Oliva et al., en 1997 descartaban la proteína 8fab por la conformación de Ramachandran de este lazo.

Sobre el dominio de proteínas cabe destacar el trabajo realizado sobre proteínas aparentemente no relacionadas, como el EGF y el PCI. Estas proteínas fueron relacionadas estructural y funcionalmente en 1998 y muchos de los datos se habían obtenido a partir de análisis preparados por inspección visual (Blanco-Aparicio et al., 1998; Mas et al., 1998). Gracias a esta metodología ha sido posible determinar más exhaustivamente las relaciones estructurales entre los esqueletos polipeptídicos de estas proteínas. Los resultados correspondientes a esta parte del trabajo se muestran en el último capítulo de tesis.

Finalmente, el método se muestra muy útil en la determinación de clases topológicas de los *folds* proteicos. Permite obtener alineamientos estructurales sin necesidad de establecer equivalencia atómica entre las proteínas y obviar el problema de la traslación previa al cálculo del RMSD, lo cual podía implicar un cierto error para proteínas no homólogas.