UNIVERSITAT AUTÒNOMA DE BARCELONA

Departament de Bioquímica

Institut de Biologia Fonamental "Vicent Villar i Palasí"

COMPARACIÓN COMPUTACIONAL DE ESTRUCTURAS DE PROTEÍNAS. APLICACIÓN AL ESTUDIO DE UN INHIBIDOR DE CARBOXIPEPTIDASA COMO AGENTE ANTITUMORAL.

JOSÉ MANUEL MAS BENAVENTE

2000

UNIVERSITAT AUTÒNOMA DE BARCELONA

Departament de Bioquímica

Institut de Biologia Fonamental "Vicent Villar i Palasí"

COMPARACIÓN COMPUTACIONAL DE ESTRUCTURAS DE PROTEÍNAS. APLICACIÓN AL ESTUDIO DE UN INHIBIDOR DE CARBOXIPEPTIDASA COMO AGENTE ANTITUMORAL.

Memoria redactada para optar al Grado de Doctor en Ciencias, sección de Bioquímica, por la Universidad Autónoma de Barcelona, por

José Manuel Mas Benavente

V.B.

Los directores de la tesis

Dr. F. Xavier Avilés i Puigvert

Dr. Enrique Querol Murillo

Bellaterra, Diciembre de 2000

a Lumi, Oscar y Esther

Prólogo

Esta tesis pretende reflejar el trabajo realizado durante los años 1995 a 1999. He intentado dar lo mejor de mi limitada capacidad de expresión y mi sentido del buen gusto en la presentación y formato de este texto. Espero no resultar más desagradable de lo estrictamente necesario. En arras de hacerme entender, he sido repetitivo y he descuidado ciertas formalidades para hacer la lectura más laxa.

Como cualquier trabajo de investigación, siempre se puede hacer más y mejor, siempre aparecen nuevas preguntas y nuevas posibles respuestas, y eso sucede justamente aquí. Revisando el trabajo, uno se pregunta el por qué de infinidad de cosas y el por qué no de otro infinidad. Preguntas que quedan abiertas para futuros doctorandos que sin duda mejorarán lo que aquí se expone, completarán lo que aquí se presenta incompleto y dejarán nuevos porqués y nuevos porqué nos... Sirva como ejemplo de lo anterior el trabajo respecto del PCI y el EGF, basado en datos no contrastados y a la espera de nuevos resultados que respondan preguntas y eliminen hipótesis. Pero no son tan sólo cuestiones formales sino, ¿alguien podría decirme qué extraño motivo me movió a hacer algunas imágenes con fondos de colores? Preguntas sin respuestas.

Abreviaciones empleadas

A lo largo de este trabajo se han empleado una serie de palabras anglosajónicas. Su significado se describe a continuación teniendo en cuenta el/los significados que han ido adquiriendo en el texto.

Coil Región de la proteína que no presenta una estructura secundaria tipo hoja beta o hélice alfa.

Core Región central de las proteínas.

Cluster Grupo de elementos pertenecientes, según algún criterio estadístico, a una población o conjunto que los contiene.

Clustering Proceso de formación de los clusters.

EGF Factor de crecimiento epidérmico.

Fold Plegamiento de la proteína. También se utiliza para referirse a la arquitectura global de la proteína.

Loop Es un lazo, una zona de la proteína enmarcada entre dos estructuras secundarias regulares. Muchas veces se utiliza para definir una región más amplia o no acotada por elementos estructuras secundarias.

PCI Inhibidor de carboxipeptidasa de patata

PDB Banco de datos de proteínas

Reshuffling Formación y ruptura sucesiva de los enlaces covalentes de las cisteínas.

RMS/RMSD Desviación media cuadrática

Scafold Es el andamiaje de la proteína. Algunas veces no del todo correctamente se se utiliza como sinónimo de *fold*.

Small Protein Proteína pequeña es la traducción literal, pero incluye proteínas de tamaño no definido generalmente con poca o ninguna estructura secundaria regular.

Scrambles Son las distintas posibles estructuras que adopta una proteína durante el proceso de *reshuffling.*

Los nombres propios de proteínas o familias de proteínas han sido transcritos tal y como se encuentran en el Banco de Datos de Proteínas de estructura conocida (PDB), es decir, en inglés. Sólo en casos excepcionales donde se emplean nombres genéricos han sido traducidos al castellano.

ÍNDICE TEMÁTICO

ÍNDICE DETALLADO

INTRODUCCIÓN

Breve introducción a los temas tratados .

CAPÍTULO I

Las información genética se almacena en un código unidimensional, en forma de secuencia específica de bases de nucleótidos. Las proteínas, de modo análogo, están constituidas por cadenas de aminoácidos que forman secuencias lineales específicas (Creighton, 1993). Tan sólo veinte elementos distintos, aminoácidos, justifican el amplio espectro de actividades metabólicas, usos estructurales, y demás funciones asignadas a ellas.

El principal objeto de análisis de este trabajo son las estructuras tridimensionales de las proteínas. Para ello se han desarrollado diversos programas que nos han permitido inferir relaciones estructurales y funcionales entre ellas y realizar una clasificación de proteínas. Algunos de los análisis realizados están siendo utilizados para el desarrollo experimental de nuevos fármacos destinados a la lucha contra el cáncer.

Este capítulo pretende servir como introducción a los temas que posteriormente van a ser tratados con mayor profundidad.

I-A PUENTES DE AZUFRE

Las **cisteínas** son los principales componentes de los puentes de azufre. Se trata de aminoácidos proteinogénicos de 121 u.m.a.s que se caracterizan por tener un grupo tiol o sulfhidrilo (-SH) en su cadena lateral. Esto le confiere propiedades fisicoquímicas que lo enmarcan como un aminoácido polar sin carga o cargado negativamente, según su estado de oxidación (Lehninger, 1983). El pKa (pH al cual varía el estado de oxido/reducción) del grupo tiol es de 8.33, relativamente cercano a pH fisiológicos. Es por tanto un grupo muy susceptible de variar su estado de protonación. El pKa del grupo carboxílico de la cisteína es de 1.71 y el del grupo amino es de 10.78 (Creighton, 1993).

El carácter polar de las cisteínas hace que tales residuos queden expuestos al solvente, generalmente polar, es decir, no suelen estar enterradas en el *core* o cuerpo central de la proteína, comúnmente más hidrofóbico. Las cisteínas

expuestas suelen tener un carácter funcional importante, dada su reactividad. Cuando el grupo tiol está desprotonado es altamente reactivo y es susceptible de formar enlaces covalentes con grupos reducidos. Las cisteínas reducidas, o protonadas, pueden enlazarse covalentemente con otros grupos oxidados de la misma o de otras cadenas polipeptídicas. Al enlace covalente de los grupos tiol, -S y –SH, de dos cisteínas se le conoce con el nombre de **puente o enlace de azufre o disulfuro**. Tras este enlace, las cisteínas pierden su polaridad presentando un cierto carácter hidrofóbico que les permite situarse en el *core* de las proteínas. Los puentes de azufre juegan un papel muy importante tanto en el proceso de plegamiento de las proteínas como en la estabilidad de su estructura nativa, generalmente la que le confiere la función. Durante este trabajo vamos a tratar tan sólo cisteínas enlazadas covalentemente con otras, es decir formando enlaces disulfuro.

En la Figura I.1 se puede ver la estructura de un puente de azufre. Se muestra la distancia que se considera máxima entre los dos átomos de azufre de las dos cisteínas que forman dicho enlace (2.8 Å).

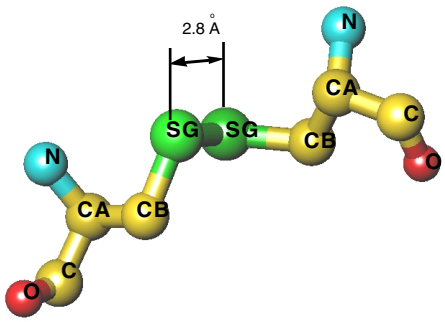


Figura I.1 Enlace o puente de azufre. Las abreviaciones de la figura corresponden a los siguientes átomos: N nitrógeno, O oxígeno, C carbono carboxílico, CA carbono α, CB carbono β, SG azufre.

Los puentes de azufre son los menos estables de los enlaces covalentes que posee una proteína, oscilando la energía de enlace entre 5 y 15 Kcal/mol en condiciones fisiológicas (Creighton, 1993). El ángulo de rotación del enlace para las geometrías más estables es cercano a los $+90^{\circ}$ o a los -90°. La

interconversión de estos dos rotámeros es rápida, mostrando una barrera de activación próxima a 7 kcal/mol (Creighton, 1993).

Los puentes de azufre se forman en el interior del retículo endoplasmático durante el proceso de plegamiento de la proteína, aunque se desconoce el mecanismo exacto. Se supone necesaria la presencia de un aceptor/donador de electrones para que el grupo tiol de la cisteína alcance su estado oxidado. El glutatión es la molécula portadora de grupos tiol más abundante que está presente en el interior del retículo endoplasmático (1-10 mM). Este motivo hace que se le asigne un importante papel en la formación de estos enlaces al intervenir en la reacción de óxido-reducción. Por otra parte, las protein-disulfuro isomerasas, también presentes en el interior de retículo, son proteínas que actúan como *foldonas*, ayudando en la correcta formación de los puentes de azufre, y por consiguiente, al correcto plegamiento de las proteínas.

Durante el proceso de plegamiento de proteínas ricas en cisteínas, ha sido descrito un rápido intercambio de los enlaces disulfuro, es decir, se describe la formación y ruptura de puentes de azufre. A este proceso se le conoce con el nombre de *reshuffling.* Fruto de estas formaciones/rupturas de enlaces disulfuro aparece un determinado porcentaje de proteína, dependiendo de su naturaleza y condiciones del medio, que no tiene la estructura nativa de esa proteína. A esas proteínas cuyos puentes de azufre no corresponden con los de la estructura nativa se les conoce con el nombre de *scrambles.* La población heterogénea de proteínas con distintos apareamientos de puentes de azufre, inestables en muchos casos, es a menudo utilizada para estudiar el camino de plegamiento de este tipo de proteínas. Es decir, el conjunto de moléculas que no tienen el plegamiento nativo pueden tender a un solo estado y son por tanto ilustradoras del camino del plegamiento. La tendencia de las moléculas desplegadas a pasar por determinados estados se justifica desde un punto de vista energético, ya que ciertos intermediarios del plegamiento presentan menor energía libre. Este efecto ha sido ampliamente descrito para varias proteínas, entre ellas el BPTI (*Bovine Pancreatic Trypsin Inhibitor*) (Creighton, 1993).

Los puentes de azufre pueden formarse entre cisteínas alejadas en la secuencia proteica o incluso, entre cadenas distintas de la misma o de otras proteínas (Figura I.2).

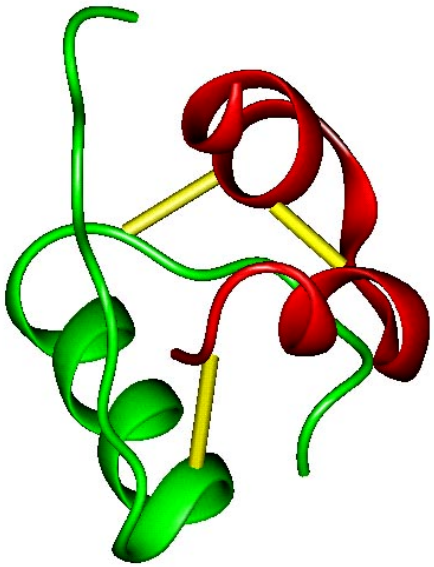


Figura I.2 Representación en cinta de dos cadenas de insulina unidas por enlaces puentes de azufre (en amarillo)

I-B CLASIFICACIÓN DE PROTEÍNAS.

Las clasificaciones estructurales de proteínas pueden ayudar en la interpretación y descripción de las relaciones estructurales y funcionales entre proteínas. El crecimiento espectacular de los bancos de datos de proteínas de estructura conocida (Figura I.3) (*Protein Data Bank*), gracias a la mejoras en las tecnologías de cristalografía de Rayos X y de Resonancia Magnética Nuclear, han hecho posible y necesario el desarrollo de las clasificaciones de proteínas. Entre las distintas clasificaciones disponibles cabe destacar tres: SCOP (Murzin A., et al. 1995), CATH (Orengo C., et al. 1996) o FSSP (Holm L. & Sander C., 1997). Las dos últimas se caracterizan por utilizar mecanismos automáticos de clasificación. La primera de las clasificaciones, SCOP, se apoya en el uso de procedimientos automáticos pero tiene en cuenta otros parámetros como la funcionalidad de la proteína, el origen evolutivo, etcétera. Para tener una visión general de cómo han evolucionado los bancos de datos se puede observar la Figura I.3.

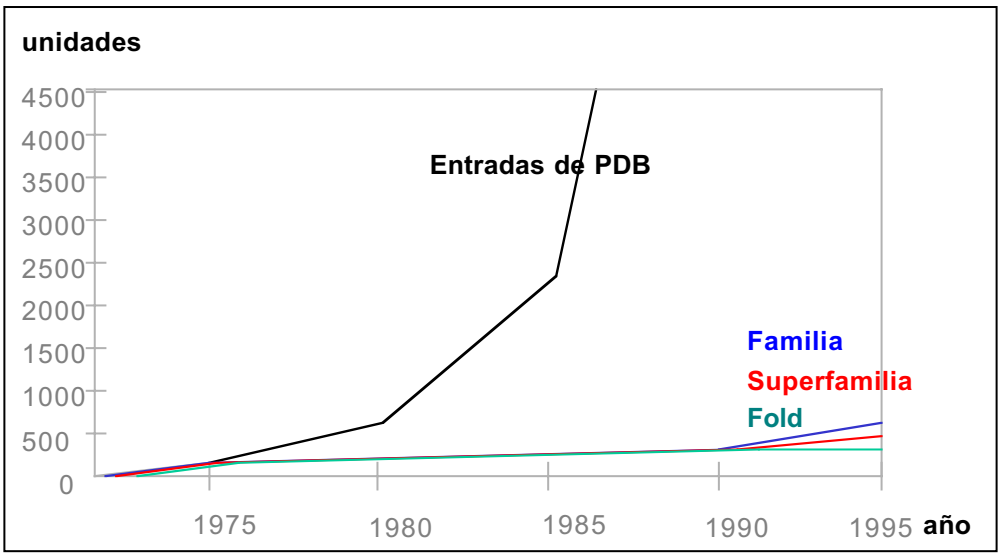


Figura I.3 El gráfico muestra el número de entradas de los últimos años en los distintos bancos de datos.

Los análisis automáticos de los bancos de datos de proteínas de estructuras conocidas son, hoy día, muy poco eficientes y ampliamente incompletos, como se comentará a lo largo de este trabajo. Los sistemas de clasificación automática permiten la clasificación del 90% de las proteínas (el resto de las proteínas no presentan los requerimientos mínimos para ser clasificadas por estas metodologías) (Orengo, 1996). Las clasificaciones que tan sólo tienen en cuenta la información analizada automáticamente desde el banco de datos incurren en contradicciones y no tienen en cuenta información de vital interés, como la función de la proteína. De este modo, y sin invalidar ninguna de las nombradas clasificaciones, muchos autores piensan que SCOP es, hoy en día, la clasificación de proteínas con mayor interés.

En general, las clasificaciones (SCOP, CATH, FSSP, etc) utilizan la información secuencial y de elementos de estructura secundaria regular para agrupar las proteínas en primera instancia. Posteriormente tienen en cuenta otros elementos, como el *fold* de la proteína, entendiendo como *fold* la arquitectura o andamiaje de las proteínas y la distribución espacial de los elementos de estructura secundaria regular. Estos factores son tenidos en cuenta en las distintas clasificaciones, las cuales dan distinto peso a la secuencia, a las estructuras secundarias regulares, etc. Este hecho hace que las clasificaciones actuales estén sometidas a una cierta subjetividad.

En la descripción de las clasificaciones de proteínas utilizadas durante este trabajo vamos a utilizar tres niveles de jerarquía descritos en SCOP y cuyos criterios coinciden ampliamente con CATH. Así, podemos definir:

- **Proteínas de similar plegamiento o *fold*:** Lo constituyen proteínas con muy alta similitud estructural. Estas proteínas tienen un andamiaje similar y los elementos de estructura secundaria regular se distribuyen de un modo similar. No tienen por qué tener el mismo origen evolutivo.

- **Superfamilia:** Proteínas con posible origen evolutivo común. Presentan muy bajo porcentaje de identidad secuencial pero con alta similitud conformacional.

- **Familia:** Implica una relación evolutiva entre las proteínas, generalmente con más del 30% de identidad secuencial. Se utilizan también criterios conformacionales y funcionales, permitiéndose en estos casos un menor porcentaje de identidad. Un ejemplo sería el caso de las globinas, las cuales tienen tan sólo un 15% de identidad en algunos de sus miembros.

Se muestra a continuación un resumen de las divisiones que realizan las clasificaciones SCOP y CATH. La clasificación que se muestra corresponde esencialmente a los criterios del SCOP y sólo cuando estos criterios son significativamente distintos en CATH, se exponen de un modo explícito.

Las proteínas se agrupan por comparación de estructuras secundarias regulares en: todo-α, todo-β, α/β, $\alpha+\beta$ y proteínas poco estructuradas (Levit & Chothia, 1976; Richarson, 1981). Se estableció la siguiente división:

Proteínas todo-α: Proteínas que presentan 3 o más hélices o bien más de un 60% hélice α y menos de un 5% de estructuras β.

***Bundle*:** Hélices α colocadas casi paralelas o antiparalelas las unas con las otras.

***Non-Bundle*:** Grupos de hélices α que no pueden ser clasificados como *bundle*. Estas hélices α suelen presentar una gran variación en ángulos y tamaños.

Poca estructura secundaria: Proteínas pequeñas, con poca estructura secundaria regular, compactas y con una o dos hélices α.

Proteínas todo-β: Proteínas ricas en hojas β ya sean paralelas, antiparalelas o mixtas (<8% en forma de hélice α y más del 45% en estructura β). Los dos grupos mayoritarios son *sandwiches* (dos hojas β retorcidas y empaquetadas) y barriles (una hoja β que va girando). Según la clasificación de CATH podemos dividir las proteínas mayoritariamente todo-β en las siguientes: *ribbon*, hoja β sencilla, *roll*, barriles, *clam*, *sandwich*, *sandwich* distorsionado, *trefoil*, prisma ortogonal, prisma alineado, 4-*propellor*, 6-*propellor*, 7-*propellor*, 8-*propellor*, 2-selenoide y 3-selenoide.

Proteínas con hélices α y hojas β: Las dos clasificaciones definen distintamente las proteínas que presentan estructuras α y β, pero siempre con más del 30% de su estructura en hélice α y más del 30% en hoja β. Así, CATH agrupa todas aquellas proteínas que presentan estas dos estructuras secundarias en un solo bloque mientras SCOP los separa en dos grupos. De este modo podemos ver las siguientes posibilidades:

para SCOP : Define dos grupos α/β y α+β, según la posición relativa de las hélices α y las hojas β.

α/β (α y β alternantes): Proteínas que presentan estas posibilidades con hojas β como elementos centrales. Así SCOP define los siguientes grupos:

- Hojas β paralelas formando un barril (con hélices empaquetadas en el exterior).
- Hojas β paralelas entregiradas (con hélices en una o ambas caras).
- Hojas β mixtas (con hélices en una o ambas caras).

α+β (α y β en disposición aleatoria): Proteínas que presentan hélices α y hojas β consecutivamente colocadas.

para CATH : Define un único grupo de proteínas y a su vez las separa en varios subgrupos: *roll*, *barrel*, 2-capa-sandwich, 3-capa-sandwich, 4-capa-sandwich, *box*, *horseshoe*, y estructuras poco estructuradas.

Proteínas con poca estructura secundaria regular o *coil*: Son proteínas que tienen una pequeña proporción de estructuras secundarias regulares o que no pueden ser asignadas a otras clases. Estas proteínas suelen utilizar puentes de azufre o iones metálicos para su estabilización. Suelen ser proteínas de pequeño tamaño.

I-C EL CÁNCER

Una célula cancerosa es, dicho de un modo superficial, una célula que se está dividiendo constantemente, o más precisamente, una célula cuya fase G_0 es muy breve. Hoy día se acepta que la acumulación de mutaciones genéticas (entre 3 y 7 independientes y generalmente implicadas en el proceso de división celular o su regulación) puede provocar la transformación de una célula en maligna o cancerosa. El cáncer parte de una sola célula que adquiere una mutación y que va adquiriendo un fenotipo más transformante e incluso invasivo en una serie de ciclos de mutación. Este proceso acaba con la aparición de células que se dividen sin control, **cáncer**.

Los **protooncogenes** son genes que codifican proteínas que intervienen en el proceso de división celular en cualquiera de sus pasos. Una mutación en estos genes, que implique una sobreexpresión o hiperactividad de este tipo de proteínas, puede inducir la proliferación descontrolada de las células, revirtiendo en un cáncer. Las mutaciones sobre este tipo de genes, **oncogenes**, suelen ser de tipo dominate. Cabe destacar como protooncogenes aquéllos que codifican factores de crecimiento, receptores de estos factores de crecimiento, los transductores de las respuestas de los receptores o los factores de transcripción que actúan como mediadores de la expresión inducida por los factores de crecimiento, es decir, cualquier proteína que intervenga en el proceso de división celular mediada por una señal externa de división (factor de crecimiento).

Una de las características más destacables de los tumores malignos es su capacidad de invasión y diseminación por otros tejidos (**metástasis**), y su capacidad para generar tumores secundarios. Los tumores de origen epitelial se conocen con el nombre de **carcinomas**. Estos tumores son los que tienen el peor pronóstico ya que suelen ser con demasiada facilidad tumores metastásicos. Entre ellos cabe destacar el de pulmón con muy alta incidencia en Cataluña, el de mama, el de próstata o el de colon-recto. Se ha descrito que estos tumores tienen mayoritariamente alterada las vías de transducción de la señal de los factores de crecimiento (Lee & Col., 1992). En este proceso el EGF (*Epidermal Growth Factor*), el TGFα (*Transforming Growth Factor* α) y el receptor del EGF

(EGFr) tienen un papel preponderante, ya que están implicados en los procesos de crecimiento tumoral, vascularización, invasividad y metástasis tumoral. Estos tumores suelen desembocar en una sobreexpresión de receptores celulares y de ligandos de la familia de los EGF, sobretodo TGFα, lo que permite una estimulación autocrina que hace muy potente este proceso.

Los factores de crecimiento

De entre los factores de crecimiento podemos distinguir un grupo que forman la familia *EGF-like*. Esta familia está integrada por más de 15 proteínas diferentes, de las cuales en este trabajo utilizamos siete (ver capítulo II) .

La mayoría de estas proteínas se sintetizan como precursores de membrana que posteriormente son procesados liberándose el factor soluble (Beil et al., 1989). Las características estructurales de estas proteínas son ampliamente tratadas en este trabajo (ver capítulos III y IV). Entre ellas cabe destacar la presencia de seis cisteínas formando tres puentes de azufre. Además tienen un hoja β y un lazo*,* conocido como lazo *B,* que coincide con la estructura de proteínas de otras familias. Esta estructura conforma un dominio que puede encontrarse en proteínas como la prostangandin sintasa (1prh), en la cual representa poco más del 10% de la proteína.

El EGF, el más representativo de la familia y el que le da nombre, puede encontrarse en los fluidos corporales como la sangre, la leche materna, la saliva o el flujo seminal. Su papel fisiológico está relacionado con la transmisión de señales que afectan a la división y la diferenciación celular. Su actividad suele implicar variaciones en el transporte iónico, en la fosforilación de proteínas endógenas, alteraciones de la morfología celular y estimulación de la síntesis de DNA (Pimentel, 1994).

Los receptores

Existen cuatro receptores diferentes descritos capaces de unirse a estas proteínas, la familia de los ErbB, llamado ErbB-1, ErbB-2, ErbB-3 y ErbB-4. Estos receptores tienen entre 1210 y 1243 residuos en humanos. Cada uno de ellos tiene un dominio extracelular rico en cisteínas, un dominio transmembrana sencillo y un gran dominio intracelular con actividad tirosinquinasa junto con varias tirosinas susceptibles de ser fosforiladas (Reise & Stern, 1998). En las formas maduras del receptor se observa una gran glicosilación, lo cual hace que aumente tanto su masa como su volumen.

Cada molécula de receptor es capaz de unirse a una molécula de ligando (1:1) permitiéndose en estos casos una homo/heterodimerización de moléculas receptor-ligando, y la posterior estabilización del dímero (Lemmon, 1997). La dimerización implica una variación estructural en el receptor que acaba con la activación de la actividad tirosinquinasa del dominio citoplasmático. Este hecho hace que se produzca una fosforilación cruzada de las tirosinas de las dos moléculas de receptor, que acaba produciendo la respuesta fisiológica anteriormente descrita.

Posteriormente estas moléculas (receptor-ligando) son internalizadas por endocitosis y degradadas, en un determinado porcentaje, en los lisosomas. La célula disminuye así las cantidades de receptores en su superficie (*down-regulation*) de tal manera que se insensibiliza a una sobreestimulación. Los distintos tipos de receptores muestran afinidades selectivas a los distintos ligandos: así los receptores ErbB-1 son esencialmente afines al EGF, mientras el ErbB-3 y ErbB-4 son mayoritariamente afines a heregulinas (otras proteínas de la misma familia).

Durante este trabajo, especialmente en el último capítulo, se analiza la familia *EGF-like*, y se profundiza en la estructura y función del EGF, cuyo papel en el proceso de acción del carcinoma es muy destacable.

OBJETIVOS

Descripción de los objetivos generales de esta tesis

CAPÍTULO II

El **objetivo general** del presente trabajo se enmarca en un proyecto más amplio de ingeniería de proteínas que trata de analizar y rediseñar la estructura, camino de plegamiento, función natural y aplicaciones biotecnológicas de una proteína, el PCI (*Potato Carboxypeptidase Inhibitor*). Las características estructurales de esta proteína, esencialmente sus puentes de azufre, nos invitaron a realizar un estudio general en proteínas ricas en puentes de azufre, cuyos resultados y conclusiones serán expuestos a lo largo de este trabajo de tesis.

Son **objetivos específicos** de esta tesis los siguientes:

1. Realización de un programa estándar para el análisis estructural de topologías disulfuro en proteínas ricas en puentes de azufre. Esto implica implementar tal programa con los algoritmos descritos en la literatura convenientes para tal fin y el diseño, creación e implementación de nuevas técnicas que permitan optimizar los resultados.

2. Completar el anterior programa con los algoritmos necesarios para obtener una clasificación automática de proteínas ricas en puentes de azufre.

3. Preparar el mismo programa para obtener superposiciones de proteínas basadas en la topología de sus puentes de azufre a fin de permitir análisis estructurales más completos.

4. Realizar un programa que permita la comparación estructural del esqueleto polipeptídico de este tipo de proteínas. Este programa deberá permitirnos analizar de un modo sencillo y flexible las estructuras de proteínas ricas en puentes de azufre.

5. Realizar un análisis estructural comparativo entre el PCI y las proteínas de mayor similitud estructural mediante el uso de las herramientas desarrolladas.

6. Determinar las claves estructurales entre el PCI y el EGF que justifican las relaciones funcionales descritas entre ambas.

Durante el desarrollo de este trabajo se han creado programas y obtenido resultados paralelos a los planteados en estos objetivos. Estos objetivos y los resultados obtenidos se detallarán a lo largo del trabajo.

METODOLOGÍA PARA EL ANÁLISIS DE LOS PUENTES DE AZUFRE

Descripción de los métodos y algoritmos implementados para el análisis de las topologías de los puentes de azufre

CAPÍTULO III

En este capítulo se describen los métodos utilizados para la comparación estructural de la topología de puentes de azufre empleados durante este trabajo de tesis.

El lector encontrará una descripción teórica y matemática detallada de cada método o algoritmo empleado. Para alcanzar algunos objetivos, como la comparación de las topologías de puentes de azufre que se describe en este capítulo, se han desarrollado más de una metodología a fin de determinar aquélla que fuera más óptima para cada caso concreto. En estos casos, todos los métodos disponibles han sido descritos y comparados, especificando en cada caso cuál ha sido el algoritmo empleado y qué motivos nos llevaron a tomar tal decisión.

El lector puede ver también en este capítulo la descripción de los algoritmos que han sido empleados para generar la clasificación de proteínas basada en la topología de sus puentes de azufre que se presenta en esta tesis.

Todos los algoritmos, tanto aquéllos empleados en la comparación estructural de puentes de azufre como aquéllos empleados para obtener unas clasificación de proteínas, han sido agrupados en un programa al que hemos llamado KNOT-MATCH. El programa KNOT-MATCH está accesible desde Internet en la dirección http://luz.uab.es. En esta dirección el usuario encontrará, además del nombrado ejecutable, una ayuda sobre el uso de este programa en varios formatos distintos.

Para facilitar la descripción y lectura de los métodos descritos en este capítulo y empleados a lo largo de todo el trabajo de tesis, los algoritmos han sido

agrupados en tres apartados según su funcionalidad. Estos apartados son los siguientes:

III-A **Sistemas para la obtención de la información estructural de los puentes de azufre.**

III-B **Métodos para la comparación estructural de puentes de azufre.**

III-C **Técnicas de clasificación o *clustering.***

III-A SISTEMAS PARA LA OBTENCIÓN DE LA INFORMACIÓN ESTRUCTURAL DE LOS PUENTES DE AZUFRE

Paso previo a todo análisis es obtener los datos que se pretenden analizar. En nuestro caso vamos a utilizar una base de datos de proteínas de estructura conocida, concretamente el *Protein Data Bank* (PDB). Este banco de datos contiene las coordenadas cartesianas de todos los átomos de estas proteínas, ya sean obtenidos mediante técnicas de difracción de Rayos X o por Resonancia Magnética Nuclear.

El principal objetivo de este apartado es disponer de la información estructural referente a la topología de los puentes de azufre del PDB de la manera más eficiente posible.

Hemos utilizado dos metodologías distintas que nos permiten disponer de la información estructural necesaria para llevar a cabo los análisis propuestos. Estas dos metodologías implican a) el uso de coordenadas cartesianas, y b) el uso de coordenadas internas de las proteínas.

Uso de coordenadas cartesianas.

Los datos estructurales disponibles en el PDB referentes a los puentes de azufre, objeto de nuestro estudio, están disponibles en forma de coordenadas cartesianas. El hecho de disponer de la información en tal formato confiere una serie de ventajas e inconvenientes que a continuación se detallan:

Ventajas:

- Éste es el formato disponible en el PDB, luego su uso implica tan sólo la recogida de un modo ordenado de la información.
- Esta forma de almacenar la información estructural permite disponer de la máxima información topológica posible.

Inconvenientes:

- El inconveniente principal para el uso de este sistema es puramente técnico. Las coordenadas cartesianas de un par de puentes de azufre (cuatro cisteínas) implican 42 bytes. Esta cantidad es, como se verá posteriormente, definitiva para la realización de determinados análisis.

Uso de coordenadas polares internas

El sistema clásico de almacenamiento de la información estructural puede ser fácilmente sustituido por coordenadas internas basado en un sistema de coordenadas polares.

Asimilamos un puente de azufre a un vector, el cual está descrito por la unión de los carbonos alfa (C_α) de las dos cisteínas que forman el enlace. Posteriormente buscamos todos los posibles pares de puentes de azufre de la proteína. Se calcula la distancia que separa el punto medio de los dos vectores (d) y el ángulo que forman dichos vectores (α). Las coordenadas de las cuatro cisteínas han sido reducidas a dos valores, **distancia** y **ángulo**. Utilizando este sistema podemos enumerar los pares de puentes de cada cadena proteica y asignar a cada uno de ellos un valor de ángulo y otro de distancia (ver Figura III.1).

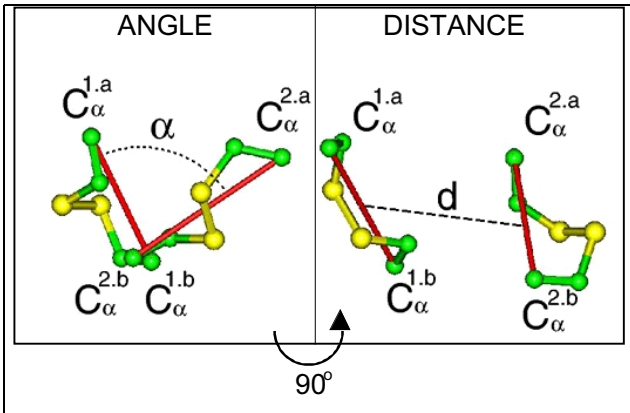


Figura III.1. Esta figura muestra gráficamente como se obtienen los valores de ángulo (α) y distancia (d). Se marcan de color amarillo los C_α de las cisteínas de dos puentes disulfuro. En color verde se muestran los átomos de azufre.

Este sistema de almacenamiento de la información topológica de los puentes de azufre plantea una serie de ventajas y de inconvenientes que se describen a continuación.

Ventajas:

- 8 bytes definen un par de puentes, seis veces menos que el método de coordenadas cartesianas
- La información recogida contiene las características estructurales necesarias para el análisis propuesto.
- La disminución de parámetros implica un aumento en la agilidad de futuros cálculos.

Inconvenientes:

- La recogida de estos valores implica un pequeño tratamiento matemático.
- Se pierde una parte de la información estructural de la proteína, aunque se trata de información prescindible para el análisis que nos interesa.

El número de pares de puentes de azufre tratado en este trabajo es demasiado grande como para tratar la topología de puentes disulfuro utilizando el sistema de coordenadas cartesianas. El sistema de coordenadas internas no implica la pérdida de la información estructural objeto de análisis. Por tanto, consideramos más adecuado este sistema para llevar a cabo el análisis propuesto. No obstante, es interesante disponer del método de análisis sin simplificaciones para poder contrastar los resultados obtenidos.

III-B MÉTODOS EMPLEADOS PARA LA COMPARACIÓN DE TOPOLOGÍAS DE PUENTES DE AZUFRE EN PROTEÍNAS

Se describen a continuación las metodologías matemáticas diseñadas e implementadas en el programa KNOT-MATCH para comparar estructuralmente la topología de puentes de azufre de dos proteínas.

El principal objetivo de este apartado es diseñar la estrategia más óptima para comparar estructuralmente topologías de puentes de azufre. Esta metodología debe cumplir requerimientos de **fiabilidad**, en el caso de utilizar aproximaciones, y **agilidad**, dado el elevado número de comparaciones que deberán realizarse.

Al comparar dos elementos cualesquiera, en nuestro caso dos topologías de puentes de azufre, precisamos de un valor numérico que describa la similitud/diferencia de las elementos que se están comparando. En biocomputación el elemento de comparación suele ser la estructura de dos proteínas y el valor obtenido de tal comparación es un valor de distancia, el RMS *(Root Mean Square)* o RMSD *(Root Mean Square Deviation)*. Resulta imprescindible comprender el significado matemático del RMSD para poder evaluar los distintos métodos diseñados.

Han sido diseñados un grupo de algoritmos para la comparación estructural de topologías de puentes de azufre de proteínas. Se han diseñado tres estrategias distintas que nos van a permitir:

- Disponer de un sistema fiable y ágil para comparar topologías de puentes de azufre,

- Disponer de un sistema de control de resultados,

- Disponer de un mecanismo que permite el análisis de proteínas evolutivamente relacionadas, es decir, proteínas donde la asignación de átomos equivalentes está descrita en la literatura o por el alineamiento secuencial.

III-B.1 Cálculo del RMSD (Desviación media cuadrática)

El RMSD es una distancia que da idea de la diferencia estructural de dos topologías. De este modo, cuanto menor sea el valor de RMSD de comparación de dos estructuras mayor es la similitud estructural de las mismas. La descripción matemática del cálculo es la siguiente:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\left(x_i - x'_i\right)^2 + \left(y_i - y'_i\right)^2 + \left(z_i - z'_i\right)^2\right]}$$

Donde N es el número de átomos que tiene el sistema y (x,y,z) son las coordenadas de los átomos de la primera estructura y (x',y',z') son las coordenadas de los átomos equivalentes[1] de la segunda estructura.

Previo al cálculo de RMSD es preciso superponer las dos estructuras que se desean comparar. El valor de RMSD corresponde al menor valor estimado tras aplicar una serie de movimientos de **rotación** y de **traslación** de la segunda estructura sobre la primera. Estos movimientos tienen como finalidad superponer las estructuras y se van repitiendo iteractivamente hasta lograr la superposición de menor RMSD. El movimiento de traslación viene definido por el vector que une los centros geométricos de las dos estructuras. El movimiento de rotación es mucho más difícil de calcular. Intuitivamente, la molécula debe rotar sobre los 3 ejes del sistema cartesiano hasta que se obtiene la mejor superposición de los átomos equivalentes descritos del siguiente modo:

Sean A y A' y B y B' respectivamente puntos desplazados en una trayectoria (átomos equivalentes), se selecciona aquella rotación que minimiza la suma de distancias $\overline{AA'}$ y $\overline{BB'}$. Estos movimientos requieren multitud de cálculos sobre las matrices que describen las coordenadas de las dos estructuras que se están superponiendo. Para agilizar los cálculos se obtienen previamente los *eigenvectors* y *eigenvalues* que describen el sistema.

[1] En dinámica molecular átomos equivalentes son un mismo átomo en dos posiciones de una trayectoria. Uno de los problemas a los que nos enfrentamos es determinar cuáles son los átomos equivalentes entre dos proteínas distintas. Este problema será tratado más profundamente con posterioridad.

Los cálculos necesarios para obtener el valor de RMSD han sido realizados mediante las subrutinas KABSCH, JACOBI y CROSS del Numerical Recipes (Press, W.H., 1988). La descripción detallada de estas rutinas y los fundamentos matemáticos de las mismas no son explicados en este trabajo dado, por un lado su complejidad, y por otro que fueron empleados sin aplicar ninguna modificación sobre la descripción original de sus autores.

III-B.2 Cálculo de todas las posibilidades

El objetivo de esta metodología es disponer de un algoritmo que no utilice ninguna simplificación en la comparación de topologías de puentes de azufre. Por tanto, este algoritmo va a servir de control para algoritmos que utilicen simplificaciones.

Esta metodología calcula todos los posibles apareamientos de topologías de puentes de azufre de cada par de proteínas y aplica el cálculo de RMSD para cada uno de ellos. Para ello se utilizan las coordenadas cartesianas de las cisteínas. El alineamiento estructural con menor valor de RMSD corresponde al mejor alineamiento posible de esas dos topologías.

Un ejemplo de los requerimientos necesarios para aplicar este algoritmo sería el siguiente:

Sean dos proteínas con tres puentes disulfuro cada una. Cada proteína tiene por tanto tres pares de puentes, 1-2, 1-3 y 2-3. El número de posibles alineamientos es el producto de las combinaciones de seis elementos tomados de tres en tres de la primera proteína por el mismo cálculo para la segunda proteína. Cabe destacar que para realizar el cálculo de RMSD, cada cisteína de la primera proteína debe ser asignada a las cisteínas de la segunda proteína, ya que no sabemos cuales son los átomos equivalentes. Este cálculo corresponde a las variaciones con repetición de dos elementos tomados de tres en tres. Este último cálculo deberá realizarse en las dos proteínas y su producto debería añadirse como producto a la operación anterior. Finalmente obtendríamos el siguiente resultado,

N = 20 * 20 * 8 * 8 = 25600 combinaciones posibles

Cada combinación requiere el cálculo de RMSD. Ésta es por tanto, una manera muy "cara" computacionalmente de comparar las topologías de puentes de azufre.

La metodología descrita anteriormente nos va a permitir disponer de un sistema que no incluye ningún tipo de simplificación. Descartamos el uso de este método debido a sus requerimientos, tanto por el tamaño del banco de datos, como por motivos de tiempo de cálculo y capacidad de máquina. Este sistema podrá ser utilizado en la comparación de un número de proteínas pequeño o como sistema de control para otras metodologías.

III-B.3 Diseño del SS-Matching Method

El principal objetivo de este método es disponer de un sistema de comparación de topologías de puentes disulfuro ágil en el cálculo, con los menores requerimientos de memoria RAM posibles, y que permita descartar comparaciones que previsiblemente vayan a dar RMSD de valores altos, evitando tiempos de cálculo excesivamente elevados.

Se utilizan las coordenadas internas. Este método deriva del algoritmo *the Basic Interatomic Distance Matching Method* descrito por Taylor & Orengo en 1989. El algoritmo original utiliza las distancias interatómicas y ha sido necesario modificarlo para adaptarlo al uso de coordenadas internas.

Se procede a la descripción matemática del algoritmo.
La matriz $\overline{\overline{S}}_{i,k}$ contiene las comparaciones de los puentes disulfuro de la proteína A con los puentes disulfuro de la proteína B. Definimos los elementos de $\overline{\overline{S}}_{i,k}$ del siguiente modo:

$$S_{i,k} = \sum_{m=1}^{N} \frac{A\_d}{\left(\left| \frac{d^{A}_{i,i+m} - d^{B}_{i,i+m}}{max\_d} \right| + B\_d \right)} + \frac{A\_\alpha}{\left(\left| \frac{\alpha^{A}_{k,k+m} - \alpha^{B}_{k,k+m}}{max\_\alpha} \right| + B\_\alpha \right)}$$

Donde $S_{i,k}$ es el valor calculado de similitud para el par de puentes disulfuro (i) de la proteína A y el par de puentes (k) de la proteína B. "d" y "a" son respectivamente los valores de las distancias y ángulos para cada puente de azufre (ver Figura III.1), siendo "max_d" (57.0 Angstroms) y "max_a" (π radianes) el máximo valor esperado[2] para ángulos y distancias. "A_d" y "A_a" permiten establecer la relación distancia/ángulo (valores calculados por iteracción 75.0 y 25.0 respectivamente). Finalmente, "B_d" y "B_a" previenen errores de división por 0, su mejor valor estimado iterativamente es 1.0.

Utilizando los valores de las constantes descritas anteriormente obtenemos valores de $S_{i,k}$ entre 0, para ninguna similitud estructural, y 100, para la identidad estructural (dos topologías idénticas). Dada la arquitectura del programa, la modificación de los valores de estas constantes por el usuario resulta muy sencilla, pudiéndose adaptar la fórmula a las necesidades de la búsqueda.

Supongamos que queremos hacer un análisis de topologías de tres puentes de azufre. Una proteína con tres puentes tendrá una sola combinación de sus puentes para distribuirse en el espacio, la cual incluirá todos sus puentes. Una proteína con cuatro puentes de azufre (1, 2, 3 y 4), analizando topologías de tres puentes de azufre, dispondrá de tres posibles topologías, correspondientes a los siguientes agrupamientos de puentes: (1, 2 y 3), (1, 2, 4) y (2, 3, y 4). Este hecho hace que la matriz $S_{i,k}$ no sea necesariamente una matriz cuadrada.

El conjunto de comparaciones individuales de topologías de puentes de azufre de dos proteínas se almacenan en matrices que contienen una compleja información: las distintas topologías de puentes de azufre que puede tener una

[2] Estos valores fueron determinados experimentalmente.

proteína frente a las topologías posibles de la proteína con la que se compara, los puentes implicados en la formación de ambas topologías y el valor de RMSD de cada comparación.

La cantidad masiva de información y cálculos necesarios para poder llevar a cabo el análisis excede con mucho las posibilidades técnicas de las que disponemos. Por tanto, es necesario definir un sistema a partir del cual sea posible evitar el máximo número de cálculos perdiendo la menor información estructural posible. Consecuentemente definimos la topología de puentes de azufre de la proteína A partiendo del conjunto de vectores descritos sobre los C_α de los disulfuros de los N puentes de azufre de la proteína. Por tanto $S^i_{A_N}$ corresponderá a la topología del i-ésimo puente de azufre de la proteína A, pudiéndose escribir como $S^{i}_{A_N} = \left\{ r^{SS(A,i)}_{C_\alpha} \right\}_{i=1,N}$. Donde $r^{SS(A,i)}_{C_\alpha}$ contiene el par de vectores que describen las posiciones de los C_α de las cisteínas que forman el puente disulfuro de orden i-ésimo. Por tanto, quedan definidos sobre $\mathbf{R^3} \times \mathbf{R^3}$. Análogamente al cálculo anterior, podemos escribir $S^{j}_{B_N} = \left\{ r^{SS(B,j)}_{C_\alpha} \right\}_{j=1,N}$ para la proteína B. Definimos ahora cRMSD como el menor RMSD para la comparación individual de dos topologías de las proteínas A y B, matemáticamente sería descrito como:

$$cRMS\left(S^{i}_{A_N}, S^{j}_{B_N}\right) = \min\left\{ x \in R;\ x = \sqrt{\frac{\sum_{i=1}^{N} \left| \overline{\overline{R'}} \left(r^{SS(A,i)}_{C_\alpha} \right) - r^{SS(B,i)}_{C_\alpha} \right|^2}{2N}} \ \forall\ \overline{\overline{R'}} \right\}$$

Definimos una aplicación $T : R^3 \rightarrow R^3$, donde la extensión de T en $\mathbf{R^3} \times \mathbf{R^3}$ viene dada por:

$$T' : R^3 \times R^3 \rightarrow R^3 \times R^3$$
$$T'(x, y) = \left(T(x), T(y)\right)$$

Por tanto, $\overline{\overline{R}}'$ es la extensión en $\mathbf{R^3} \times \mathbf{R^3}$ de una rotación $\overline{\overline{R}}$ en R^3. Definimos ahora $\Im_N^{A,B}$ como el conjunto de comparaciones de todas las topologías de puentes de azufre de las proteínas A y B. Este conjunto se puede definir como:

$$\Im_N^{A,B} = \left\{ x;\ x = \text{cRMSD}\left(S_{A_N}^i, S_{B_N}^j\right) \forall \text{ combinaciones en los conjuntos } S_{A_N}^i \text{ y } S_{B_N}^j \right\}$$

El conjunto de combinaciones que tiene en cuenta todas las posibilidades, incluidas las orientaciones paralela-antiparalela de los puentes, crece exponencialmente al aumentar N. De este modo, ha sido necesario definir un sistema de discriminación que permita eliminar elementos de este conjunto. Así, definimos un conjunto reducido $\overline{\Im}_N^{A,B} \leq \Im_N^{A,B}$. Este conjunto se describe como:

$$\overline{\Im}_N^{A,B} = \left\{ x;\ x \in \Im_N^{A,B}, \text{ de modo que } \forall\ y \in \Im_N^{A,B} \Rightarrow x \leq y \text{ y } \text{Card}\ \overline{\Im}_N^{A,B} \leq 30 \right\}$$

Siendo $\text{Card}\ \overline{\Im}_N^{A,B}$ el total de elementos en $\overline{\overline{\Im}}_N^{A,B}$. Por tanto, $\overline{\Im}_N^{A,B}$ puede contener hasta los 30 mejores alineamientos de las topologías de los dos puentes de azufre.

En resumen, no todas las comparaciones individuales de topologías serán consideradas. Sólo en el caso que el producto del número de puentes de azufre de ambas proteínas sea inferior a 30, serán consideradas todas las posibles combinaciones. En caso contrario, serán seleccionadas para este cálculo aquellas 30 topologías que presenten el menor cRMSD, según los criterios anteriormente expuestos.

Los controles realizados sobre esta metodología permiten observar que se eliminan un importante número de comparaciones de puentes de azufre. Hemos comprobado cómo las comparaciones eliminadas corresponden a altos valores de ssRMSD. No obstante, en proteínas con elevado número de puentes y

distribución isotrópica de los mismos, la simplificaciones pueden implicar un cierto error. Este error puede minimizarse variando las condiciones de ejecución del programa. El "límite de resolución" de la técnica dependerá de la capacidad de la máquina, el tiempo de cálculo y las condiciones de ejecución. Las condiciones de ejecución del programa que van a ser utilizadas minimizan los errores y no afectan a los cálculos y conclusiones finales.

III-B.4 Alineamientos estructurales de un modo secuencial (Dynamic Algorithm)

Este algoritmo fue descrito en 1989 por Taylor y Orengo y es un método clásico para el alineamiento de secuencias de proteínas y DNA. El objeto de introducir este algoritmo es utilizar un sistema similar para el alineamiento estructural. Este método nos permitirá saber si el orden secuencial en las cisteínas que forman los puentes de azufre implican la conformación topológica de estos disulfuros.

Este algoritmo implica el alineamiento de elementos de un modo secuencial. En nuestro caso el alineamiento tiene un carácter estructural, por tanto, establecemos como orden secuencial el orden de aparición de las cisteínas implicadas en los puentes de azufre.

La matriz de valores con la información estructural de la proteína es introducida en el *Dynamic Algorithm* a través de la matriz $\overline{\overline{S}}$ o matriz de escores, definidas en *SS Matching Method* (Taylor & Orengo, 1989).

El algoritmo se puede describir matemáticamente del siguiente modo:

Sea $\overline{\theta}_x^{\overline{\overline{S}}}(p,q)$ una función vectorial aplicada sobre la matriz $\overline{\overline{S}}$, donde p y q son respectivamente fila y columna de "inicio de búsqueda". $\overline{\theta}_x^{\overline{\overline{S}}}(p,q)$ se describe como:

$$\bar{\theta}_x^{\bar{\bar{S}}}(p,q)\left\{\begin{array}{l}\theta_1^{\bar{\bar{S}}}(p,q) = S_{k=p+1,l=q+1} \\ \theta_2^{\bar{\bar{S}}}(p,q) = \max_{k=i+2\rightarrow npa} S_{k,l=j+1} - g - g' \\ \theta_3^{\bar{\bar{S}}}(p,q) = \max_{l=j+2\rightarrow npb} S_{k=i+1,l} - g - g'\end{array}\right\}$$

donde "npa" y "npb" son respectivamente el número de pares de puentes de las proteínas A y B. Este número de pares de puentes corresponde a las combinaciones del número de puentes disulfuro de una proteína tomados de dos en dos. "g" y "g´" (penalización por *gap* y penalización por tamaño del *gap*) representan valores de penalización para favorecer que no se produzcan *gaps* en los alineamientos. El valor de "g" es un valor constante mientras g' debe relacionarse con el tamaño del gap. Estos valores dependen de las condiciones en las que el usuario ejecuta el programa.

Definimos la función $\left|\theta'\right|_{max}$ aplicada sobre $\bar{\theta}$ de modo que busca el valor máximo entre las componentes del vector, cuya descripción matemática es :

$$\left|\theta'\right|_{max} = \max\left\{\theta_1, \theta_2, \theta_3\right\}$$

Definimos la función ψ tal que $\psi(\bar{\bar{S}}) = \overline{\overline{SDA}}$. Así, $\overline{\overline{SDA}}$ se construye del siguiente modo:

$$SDA_{i,j} = S_{i,j} + \left|\theta'(i,j)\right|_{max}$$

Definimos ahora (i_{max},j_{max}) de modo que corresponden a los valores de (i,j) que contienen el máximo valor de $\overline{\overline{SDA}}$, es decir, el valor del alineamiento será $SDA_{imax,\,jmax}$ y el mejor alineamiento corresponderá al recorrido realizado por la función $\psi(\bar{\bar{S}})$. Esquemáticamente se puede representar el algoritmo tal y como se muestra en la Figura III.2:

$$\overline{\overline{S}} = \begin{pmatrix} 100 & 10 & 10 \\ 10 & 100 & 0 \\ 10 & 10 & 100 \end{pmatrix} \xrightarrow{\varphi(\overline{\overline{S}})} \overline{\overline{SDA}} = \begin{pmatrix} 300 & 10 & 10 \\ 20 & 200 & 0 \\ 10 & 10 & 100 \end{pmatrix} \Rightarrow score = 600$$

Figura III.2: La matriz $\overline{\overline{S}}$ muestra los valores de las comparaciones individuales de los elementos de las dos secuencias. Una secuencia estaría dispuesta de modo horizontal y la otra estaría situada verticalmente en la matriz. Tras aplicar $\psi(\overline{\overline{S}})$ obtenemos $\overline{\overline{SDA}}$, que contiene el valor del alineamiento. El recorrido indicado con flechas sobre $\overline{\overline{SDA}}$ indica el mejor alineamiento.

Así pues, disponemos de un algoritmo que permite el alineamiento estructural de proteínas utilizando los criterios clásicos del alineamiento secuencial. Este algoritmo nos permite alinear las topologías de puentes de azufre de proteínas que tengan una clara relación estructural, ya que tal método establece como átomos equivalentes las cisteínas según su orden secuencial. Este algoritmo no garantiza el mejor alineamiento de topologías de puentes de azufre entre proteínas no homólogas.

III-C MÉTODOS EMPLEADOS PARA LA CLASIFICACIÓN DE PROTEÍNAS.

En este capítulo se han descrito el conjunto de algoritmos implementados para obtener la información estructural del banco de datos y cómo proceder a la comparación de las proteínas. Las clasificaciones, incluidas las de proteínas, requieren una comparación por parejas de los elementos que se desean clasificar, aunque algunas metodologías permitan mediante simplificaciones, evitar realizar el total de las comparaciones. Procedemos por tanto, a la descripción de los métodos estadísticos que permitan obtener conclusiones de carácter general del banco de datos.

El principal objetivo de este apartado es detallar los métodos estadísticos utilizados para obtener la clasificación de proteínas basada en la topología de puentes de azufre.

Las técnicas mas comúnmente utilizadas para separar un conjunto de datos en subconjuntos naturales son las llamadas **técnicas de agrupamiento o clustering**. Hay una amplia gama de metodologías y cada una ellas se adapta mejor a unos determinados análisis y tipos de datos. De este modo podemos definir 5 grandes conjuntos de metodologías (Everit, 1974):

- Técnicas jerárquicas(HT): Permiten la construcción de dendogramas o árboles donde se relacionan los datos.
- Técnicas de optimización y particionamiento: Las agrupaciones son mutuamente exclusivas y se forman por optimización del *clustering criterion*.
- Técnicas de búsqueda de densidad (DST): Los grupos se forman por localización de regiones con alta densidad o concentración de datos.
- Técnicas de agrupación y solapamiento: Permiten solapamientos entre las distintas clases.
- Otros: Técnicas que son difíciles de encuadrar en alguno de los grupos anteriores.

Podemos definir ***cluster*** como un subgrupo de elementos pertenecientes, según un criterio estadístico, a una población o conjunto que los contiene. El principal problema se centra en la subjetividad que puede existir a la hora de optar por uno u otro método de agrupación. De este modo, los *clusters* que se forman pueden depender del algoritmo que se utilice, del mismo modo que puede depender de los parámetros o límites que introduzca el usuario. Así se dice que los *clusters* pueden llegar a ser agrupaciones relativamente subjetivas. Para evitar este fenómeno muchos autores aplican más de una técnica de *clustering* consecutivamente. Cuando los datos no forman grupos claramente aislados, sino que forman un continuo (*chaining*), realizar la separación de subgrupos o *clusters* resulta tremendamente complicado. Un ejemplo del problema expuesto sería intentar clasificar vasos de agua según estén llenos o vacíos. Si no tienen agua o si están llenos su clasificación es sencilla. Si por el contrario tenemos algún vaso con agua por la mitad su clasificación se complica y el resultado de su clasificación podría depender de los criterios o métodos que se utilicen.

Han sido seleccionadas una técnica de búsqueda de densidad o DST y una HT o técnica jerárquica. Se ha descrito que el uso de más de una técnica de *clustering* aplicada consecutivamente reduce considerablemente la subjetividad intrínseca del análisis estadístico.

III-C.1 Técnica de búsqueda de densidad (DST)

De entre la amplia oferta de técnicas de búsqueda de densidad hemos seleccionado el *Taxmap Method*. El método *Taxmap* fue descrito por primera vez en 1968 por Charmichael y colaboradores (Everit, 1974) y ha sido posteriormente tratado y modificado por otros autores. Una de las características esenciales de este método es que tiene especialmente en cuenta el problema del *chaining,* al cual se puso especial hincapié en evitar. Se caracteriza porque el usuario debe introducir un valor de *cut-off* o corte, lo que da una cierta subjetividad a los resultados y a su vez permite adaptar la técnica al tipo de datos y análisis que se desea. Esta técnica utiliza el concepto natural de agrupación, de este modo, tiende a controlar el desplazamiento del centrómero, centro de masas o geométrico, por la adición de un nuevo miembro al seno del

cluster. La descripción matemática adaptada al cálculo de nuestro interés sería la siguiente:

Sea $\overline{\overline{ssRMSD}}$ la matriz contenedora de los ssRMSD que relacionan todos los *clusters*. Definimos $\overline{\overline{R}}$ como aquella matriz que contiene todas las relaciones entre los miembros del *cluster*, por tanto

$$R_{i,j} = ssRMSD_{k,l} \ \forall \ k,l \text{ elementos cluster}$$

De un modo análogo se define $\overline{\overline{R}}'$, la cual incluye los elementos del *cluster* y al candidato a nuevo miembro del *cluster*. Definimos "d" como el desplazamiento del centrómero de posición (c) sobre la posición (c') que ocuparía si se aceptara al candidato como miembro del *cluster*. "d" sería entonces calculado como:

$$d = \left|\overline{cc'}\right| = \left(\frac{1}{C_{n,2}}\sum_{\substack{i=1\\j=1}}^{\substack{i=n\\j=n}} R_{i,j}\right) - \left(\frac{1}{C_{n+1,2}}\sum_{\substack{i=1\\j=1}}^{\substack{i=n\\j=n}} R'_{i,j}\right)$$

Donde "n" es el número de proteínas incluidas en el *cluster* y $C_{n,2}$ son las combinaciones de "n" elementos tomados de 2 en 2. Así, si el candidato a miembro del *cluster* no provoca un Δd mayor que un determinado límite, *cut-off* que controla el usuario, ese candidato pasa a formar parte del *cluster*. El centrómero del *cluster* será entonces recalculado teniendo en cuenta la presencia del nuevo miembro (c=c´), y el proceso se repite sucesivamente hasta que no aparece ningún elemento que cumpla estas condiciones. En ese momento, el *cluster* se considerará cerrado y busca la existencia de uno nuevo. Los *clusters* se inicializarán buscando de entre los elementos no enlazados aquéllos que están más próximos. Siguiendo estas reglas intuitivas hemos creado un conjunto de subrutinas que realicen este trabajo. Las matrices de datos de entrada llevan una compleja información:

- Los valores de RMSD de la comparación de las topologías de puentes de azufre de cada par de proteínas (cRMSD).
- Puentes de la proteína implicados en esa topología.
- Si la proteína ya ha sido incluida en algún otro *cluster*.

Todo ello confiere un complicado sistema de comprobaciones y cálculos que hacen que el KNOT-MATCH requiera un cierto tiempo de ejecución (16-20 horas para el análisis realizado en este trabajo). El conjunto de topologías de puentes obtenidas marcan un nuevo valor de ssRMSD.

Hemos observado cómo el aumento de la restricción de movimiento del centrómero hace que el número de *clusters* que obtenemos aumente sensiblemente. Por el contrario, una mayor tolerancia al desplazamiento del centrómero provoca la aparición del *chaining*, observándose un aumento del ssRMSD de los *clusters*. Para evitar esta degeneración, el usuario puede introducir un valor máximo de ssRMSD para un *cluster*, a partir del cual dicho *cluster* se cerrará artificialmente. No obstante, es conveniente no limitar este parámetro y utilizar condiciones más restrictivas en el desplazamiento del centrómero. Otro parámetro que debe introducir el usuario es el número mínimo de topologías que forman un *cluster*. De este modo evitamos que todo apareamiento de dos proteínas sea un *cluster*, lo que distorsionaría los resultados. El diseño del programa hace que el menor número posible de proteínas que forman un cluster sea 3. La Figura III.3 muestra gráficamente el funcionamiento del algoritmo.

Para no obviar ninguna proteína de tratamientos y estudios posteriores, aquellas que no hayan sido consideradas miembros de ningún *cluster* serán consideradas *clusters* ellas solas, siendo tratadas como tal en la segunda técnica de agrupación.

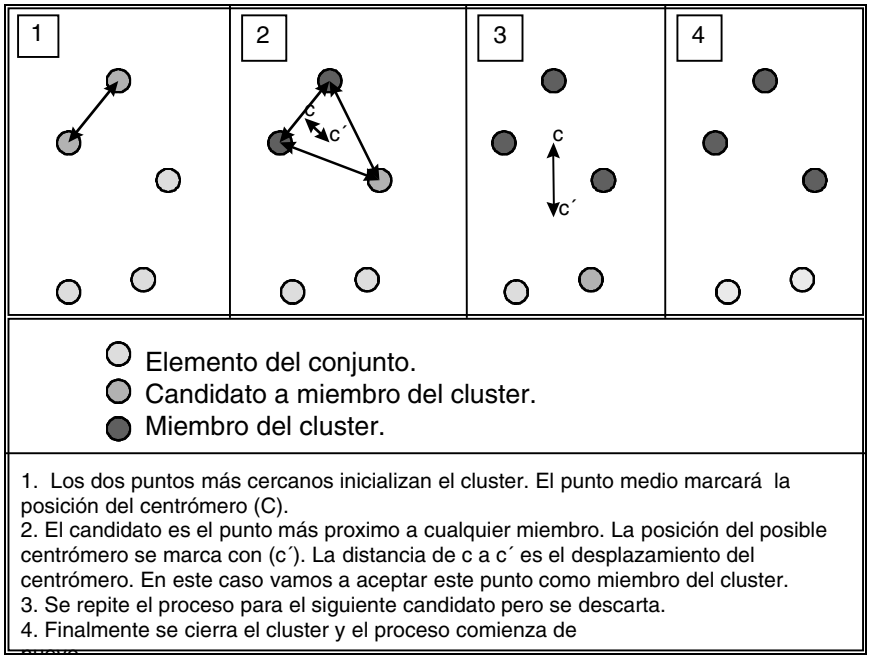
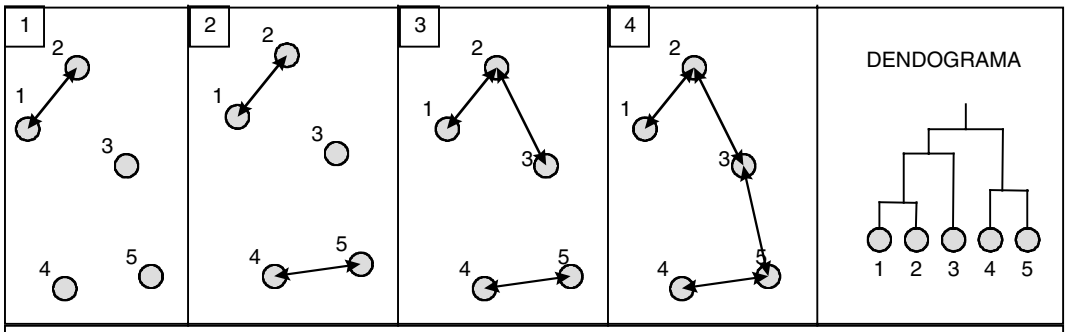


Figura III.3 Esquemas de funcionamiento de la técnica de búsqueda de densidades.

III-C.2 Técnica jerárquica (HT)

Las técnicas jerárquicas se pueden subdividir en **aglomerativas**, cuando proceden realizando una serie de fusiones sucesivas de los elementos, y **divisivas**, las cuales realizan particiones sucesivas del conjunto inicial. La idea que persigue el *Single Linkage Method*, el método que utilizamos, es agrupar en primer lugar aquellos elementos que son más parecidos. Realizada esta unión se buscan los siguientes más parecidos. Este proceso se repite hasta que todos han quedado unidos. De este modo, las primeras uniones indican una mayor similitud que las últimas. El hecho de que las primeras uniones indiquen una mayor similitud le da nombre a la metodología, *Nearest Neighbour*. La matriz de entrada de esta segunda técnica son los ssRMSD que existen entre los *clusters* o agrupaciones obtenidas mediante la DST. De este modo, se define una matriz cuadrada donde cada elemento es uno de los *clusters* que han resultado de aplicar la técnica de *clustering* anterior.

Finalmente definimos como **clases** aquellas agrupaciones de *clusters* cuyo ssRMSD sea inferior a un límite arbritario. Estas clases podrán ser divididas a su vez en subclases cuando el dendograma resultante de la ejecución de este algoritmo así lo requiera. La Figura III.4 muestra gráficamente el funcionamiento del algoritmo.



Las figuras del 1 al 4 muestran la unión secuencial de los puntos según la proximidad. Finalmente es posible apreciar el dendograma resultante de esta agrupación.

Figura III.4 Esquemas de funcionamiento de la técnica de jerárquica.

La altura de las barras verticales corresponde al cálculo siguiente:

$$a = \left(\frac{1}{C_{n,2}} \sum_{\substack{i=1 \\ j=1}}^{\substack{i=n \\ j=n}} ssRMS_{i,j} \right)$$

Siendo "i" y "j" elementos incluidos en esa rama del dendrograma y n el número total de proteínas incluidas en la misma rama.