SIMILITUDES ESTRUCTURALES MÁS ALLÁ DE LOS PUENTES DE AZUFRE

Uso de la topología de puentes de azufre de las proteínas para su clasificación y la determinación de posibles funciones heterólogas

CAPÍTULO IV

En este capítulo se describen los resultados obtenidos mediante la aplicación de los métodos descritos en el capítulo III. De este modo, cuando el lector considere insuficiente la descripción de los métodos empleados deberá consultar el capítulo III.

El capítulo ha sido dividido en dos partes que corresponden primero, a la descripción de los puentes de azufre como elementos estructurales comunes a un determinado tipo de proteínas y, en segundo lugar, a la clasificación de proteínas basada en estos elementos estructurales.

IV-A CLASIFICACIÓN DE PROTEÍNAS RICAS EN PUENTES DE AZUFRE TRAS LA OBTENCIÓN DE MOTIVOS ESTRUCTURALES COMUNES.

El programa KNOT-MATCH ha sido preparado para poder realizar una clasificación de proteínas ricas en puentes de azufre basada en la topología que describen estos enlaces en las proteínas. Para ello se utilizan las técnicas que comparan topologías de puentes entre proteínas y técnicas para obtener la clasificación propiamente dicha (ver descripción en el capítulo anterior).

IV-A.1 Introducción

El análisis en los bancos de datos de secuencias, tanto de proteínas como de DNA se utiliza frecuentemente para inferir sobre la estructura y función de proteínas, modelarlas a partir de otras de estructura conocida, etc. Para poder realizar estas tareas, hoy tan frecuentes, fue necesario desarrollar algoritmos y programas que permitieran obtener rápidos y correctos alineamientos secuenciales (para ver una revisión, Bork et al., 1994). El número de proteínas de estructura conocida aumenta constantemente, llegándose a disponer hoy día de un banco de datos con casi 10,000 proteínas cuyo nombre es *Protein Data Bank (*PDB) (Bernstein, 1977). Este crecimiento crea la necesidad de desarrollar programas que permitan realizar búsquedas específicas y clasificar la información contenida. Con afán de cubrir estas necesidades aparecieron las clasificaciones de proteínas entre las que cabe destacar: SCOP, CATH y FSSP (Murzing et al., 1995; Orengo et al, 1997; Holm & Sander 1996). Relacionar el elevado número de estructuras disponibles en el PDB no es tarea fácil. Las clasificaciones de proteínas no utilizan toda la información disponible en el banco de datos, más bien utilizan una mínima parte de ella, ya que se basan principalmente en la información secuencial de las proteínas y en el análisis de sus estructuras secundarias. Este hecho hace que sea relativamente sencillo emparentar proteínas que tienen un origen evolutivo

común, pero implica una gran dificultad a la hora de clasificar proteínas que, teniendo una estructura y/o función similar, han sufrido una evolución convergente. Cada vez es más frecuente en la literatura ver descritas relaciones estructurales y/o funcionales entre proteínas que difícilmente podrían relacionarse por los métodos secuenciales clásicos (Madej et al., 1995; Mas et al, 1998). De este modo, el análisis estructural de proteínas está falto de herramientas que permitan extraer del PDB toda la información que contiene.

Estudiar tridimensionalmente proteínas ricas en estructuras secundarias regulares y de gran tamaño es una tarea ardua que está lejos de ser hoy un proceso completamente automático. En el PDB es posible encontrar proteínas pequeñas, con poca o ninguna estructura secundaria regular, cuya secuencia no tiene homología secuencial suficiente con otras proteínas de estructura conocida. En estos casos el análisis estructural y la clasificación de estas proteínas se complica enormemente, llegando a ser una trabajo c*uasi* subjetivo. Resulta excesivamente común entre estas proteínas encontrar incoherencias como las observadas para el PCI, clasificado como una *Small Protein* o como una proteína todo β por las dos clasificaciones de proteínas más utilizadas, SCOP y CATH. Estas proteinas, pequeñas y con poca o ninguna estructura secundaria regular, suelen ser ricas en puentes de azufre, los cuales son fundamentales para mantener su estructura nativa (Chang et al, 1994 & 1995). Las funciones de estas proteínas suelen ser extracelulares siendo hormonas, feromonas, factores de crecimiento, inhibidores de proteasas, venenos, toxinas, etc... Sus funciones biológicas son, por tanto, de gran interés biomédico y biotecnológico. Así, son un grupo muy interesante de proteínas sobre las que trabajar, tanto por lo poco ajustables que son las técnicas clásicas de clasificación sobre ellas como por el interés aplicado que tienen.

IV-A.2 Objetivos

El objetivo más destacable de este apartado es establecer una clasificación de proteínas basada en la topología de sus puentes de azufre. A partir de esta clasificación se tratará de determinar los motivos estructurales basados en topologías comunes de los puentes disulfuro en el PDB. Estos motivos quedan generalmente ocultos a los métodos clásicos de análisis secuencial y estructural de proteínas. Finalmente, el análisis de las proteínas agrupadas nos permitirá establecer relaciones estructurales y/o funcionales entre algunas de ellas que hasta ahora no se han descrito.

IV-A.3 Material y métodos

El programa KNOT-MATCH es la herramienta diseñada para realizar el análisis propuesto en los objetivos de este apartado. Este programa, como ya ha sido comentado anteriormente, ha sido preparado para ser ejecutado en máquinas Silicon Graphics con sistemas operativos IRIX 5.3 y superiores. Disponemos además del banco de datos de proteínas de estructura conocida, PDB, versión 76 datada del año 1996.

Selección de las proteínas que formarán parte del sistema de análisis

El sistema donde se ha realizado el estudio se compone del conjunto de proteínas que forman el PDB. Las coordenadas atómicas de estas proteínas son conocidas, ya sea mediante técnicas de difracción de Rayos X o bien por el uso de Resonancia Magnética Nuclear (RMN). El número de estructuras incluidas en este banco es de casi 10.000. La reducción del sistema se ha realizado para hacer el estudio viable y a su vez para evitar alteraciones de los resultados debido a la elevada redundancia de algunas estructuras. De este modo, se han utilizado dos criterios que permiten hacer esta reducción:

i) Eliminación de proteínas con homología superior al 65%: Se seleccionaron los PDBs correspondientes a proteínas con un 65% o menos de homología (Hobhom & Sander, 1996). Los criterios de eliminación de proteínas descritos por estos autores son los siguientes:

- proteínas resueltas con una resolución inferior a 3.5 Angstroms.
- proteínas resueltas con R-factor superior al 30%.
- cadenas de longitud inferior a 30 residuos.
- cadenas con un número de aminoácidos no estándar superior al 5%.
- proteínas cuya homología secuencial superior al 65%.

Nota: Para obtener el sistema reducido con el que trabajar, hemos recurrido a la dirección electrónica del EMBL (European Molecular Biology Laboratory). El archivo que hemos utilizado para la realización de este trabajo data del 3 de Junio del año 1996 (ftp.embl-heidelberg.de (192.54.41.33)).

ii) Eliminación de proteínas con menos de 3 puentes disulfuro: El estudio que nosotros pretendemos realizar es sólo de aquellas proteínas que presenten tres o más puentes de azufre. Esto provoca una reducción drástica del número de proteínas que van a formar parte de nuestro estudio, llegando a disponer de un total de 159 proteínas.

A pesar de haber realizado esta importante reducción, los criterios utilizados ofrecen garantía de que no hemos perdido información estructural necesaria para la consecución de nuestros objetivos.

Condiciones de ejecución del KNOT-MATCH

Como ya ha sido descrito, el programa KNOT-MATCH dispone de varias técnicas de comparación y dos técnicas de agrupación o clasificación de proteínas. Para obtener la clasificación de proteínas que presentamos a continuación se ha ejecutado el programa KNOT-MATCH en las condiciones que describe la Tabla IV.1.

CONDICIONES DE EJECUCIÓN DEL KNOT-MATCH							
Archivo de entrada						*pdb_select (29.05.96)*	
Distancia Máxima Entre C_α en un Puente Disulfuro						*2.8 Angstroms*	
Número Mínimo de Puentes Disulfuro en una Proteína						*3*	
Desplazamiento Máximo del Centrómero (DST)						*0.3 Angstroms*	
Valor Máximo de ssRMSD para un Cluster (DST)						*0.0*	
Número Mínimo de Topologías en un Cluster (DST)						*3*	
SS MATCHING METHOD							
A_a	B_a	C_a	max_a	A_d	B_d	C_d	max_d
1.0	1.0	1.0	π	1.0	1.0	1.0	57.0
Peso de la Distancia entre el Par de Puentes						75.0%	
Peso del Ángulo del Par de Puentes						25.0%	

Tabla IV.1

Se ha utilizado el SS-Matching Method como algoritmo de comparación de topologías de proteínas y las técnicas de densidad (*Taxmap method*) y jerárquicas (*Nearest Neighbour*) descritas en el capítulo anterior.

IV-A.4 Resultados

Hemos realizado una clasificación de proteínas utilizando las dos técnicas de *clustering* descritas anteriormente. La técnica de densidad ha sido utilizada, como se comentó, para agrupar las proteínas en ***clusters***. Las proteínas agrupadas por esta técnica presentan una alta equivalencia topolóqica de sus puentes de azufre, mostrando valores bajos de ssRMSD. La técnica jerárquica ha permitido agrupar los ***clusters*** (que se ven como números en las figuras) en ***clases*** (que aparecen marcadas con letras en las figuras) que muestran significado estructural y funcional (información extraída del SCOP, del CATH y de los archivos del PDB). Los resultados se muestran en la Tabla IV.2 y en la Figura IV.1.

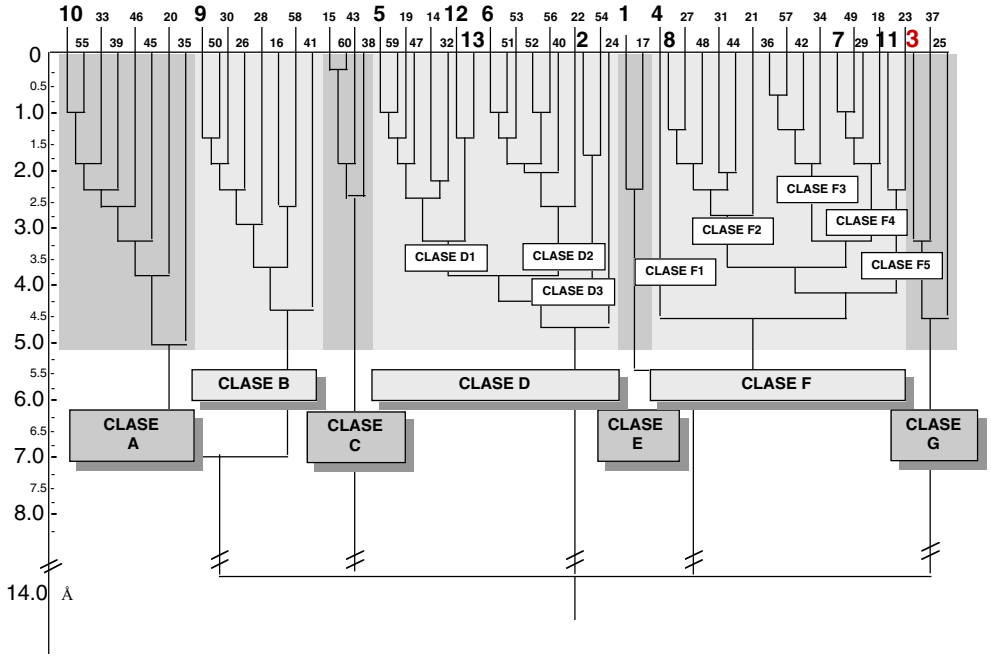


Figura IV.1. Esquema de la clasificación. Dendrograma correspondiente a la clasificación final. Aparecen recuadradas las clases y subclases obtenidas. La regla de la izquierda indica los valores de ssRMSD. En la parte superior se muestran los números correspondientes a los ***clusters*** obtenidos por la DST, primera técnica de *clustering*. Con un tamaño mayor han sido remarcados aquellos clusters que incluyen más de una proteína, tal y como se explica en el texto (ver Tabla IV.2).

En la Tabla IV.2 se puede observar qué proteína corresponde a cada uno de los números (***clusters***) y qué proteínas forman cada ***clase*** y ***subclase*** (divisiones de una clase).

Han sido obtenidos sesenta *clusters*, quince de ellos contienen más de una proteína. Once de estos quince están compuestos por proteínas con el mismo tipo de plegamiento (SCOP & CATH). Por inspección visual del dendrograma (Figura IV.1) resultante de la técnica jerárquica podemos distinguir siete clases diferentes, de la A a la G. Estas clases se ramifican en varios grupos, según la equivalencia topológica de sus puentes de azufre, es decir, cuya topología de puentes de azufre es más parecida, y por tanto, mostrando menor ssRMSD. La definición de **clases** se toma teniendo en cuenta ssRMSD inferiores a 5 Å. El criterio utilizado es arbitrario y ha sido utilizado este valor porque facilita la explicación del dendrograma. Las **subclases**, o divisiones de estas clases se obtienen separando los grupos que tengan mayor significado biológico (funcional y/o estructural) pero respetando lógicamente la estructura del dendrograma. Por

tanto, presentamos algunas de las características más llamativas de las proteínas incluidas en las distintas clases.

Tabla IV.2 : La primera columna muestra las ***clases*** y ***subclases*** en las que se ha dividido el conjunto inicial. La segunda columna tiene los ***clusters*** que forman cada ***clase***. En algunos de ellos, los compuestos por más de una proteína, se puede apreciar entre paréntesis el valor de ssRMSD de ese cluster. La tercera columna contiene la el porcentaje de clase estructural de cada ***cluster***. Finalmente, las tres últimas columnas contienen respectivamente el tipo de plegamiento, familia y código de PDB.

Cla	Cluster	Clase Estruct.	Tipo de Plegamiento	Familia	Código de PDB
	10 (0.68)	100% α	Acid Proteases	Pepsin Like	1hrnA 3psg 3cms
A	55	β	Acid Proteases	Pepsin Like	1htrB
	33	α	Hemodependent Peroxidase	Myeloperoxidase Like	1mhlC
	39	α/β	α/β Hydrolase	Fungal Lipase	1tca
	45	α/β	TIM Barrel	α-Amylase	6taa
	46	α/β	α/β Hydrolase	AcetilColinesterase	1ack
	20	α+β	Lysozime Like	Barley Endochitinase	1cnsA
	35	α/β	α/β Hydrolase	Pancreatic Lipase	1ppi
B	9 (0.53)	100% α/β	Periplasmatic Binding Protein Like II	Transferin	1lct 1nnt 1tfd
	50	α/β	α/β Hydrolase	Fungal Lipase	1tib
	30	α	4 Helical Cytokines	Long Chain Cytokines	1lki
	26	β	Segment RNA Genome Viruses Prot.	Hemagglutinin Head-Piece	1hgeA
	28	α/β	α/β–TIM Barrel	Type II Chitinase	1hvm
	16	β	Cupredoxins	Multidomain Cupredoxins	1aozA
	58	β	Inmunoglobulin Like β–Sandwinch	C_1 Set Domain	2fbjH
	41	β	Inmunoglobulin Like β–Sandwich	I Set Domain	1vcaA
C	15	β	β–Trefoil	Plant Citotoxin B Chain	1abrB
	60	β	β–Trefoil	Plant Cytotoxin B-Chain	2aaiB
	43	α	Glycosil Transferase of Superhelical fold	Glucoamylase	3gly
	38	α	4 Helical Cytokines	Short Chain Cytokines	1rcb

Continúa la Tabla IV.2

Cla	Cluster	Clase Estruct.	Tipo de Plegamiento	Familia	Código de PDB
	5 (0.37)	100%β	Trypsin Like Serin Protease	Eucariotic Protease	1hylA 1sgt 3gctA 1try
	59	β	Trypsin Like Serin Proteases	Eukariotic Protease	2kaiB
	19	β	Barwin-Like Endoglucanase	Barwin	1bw4
	47	SP	Ascaris Trypsin Inhibitor	Ascaris Trypsin Inhibitor	1ate
			Trypsin Like Serin Proteases	Eukariotic Protease	2alp
	14 (1.6)	66.6% α 33.3% β	6 Bladed β–Propeller	Sialidases	1nscA
D1			Phospholipase A_2	Vertebrate Phospholipase A_2	1poa
	32	α	Bifunctinal Inhibitor 2BIP	Bifunctinal Inhibitor 2BIP	1lpt
	12 (1.1)	100% SP	Ovomucoid PCI-1 Like Inhibitor	Animal Kazal-type Inhibitors	3sgbI 1pce 2bus
	13 (1.4)	66.6% SP 33.3% β	Ovomucoid PCI-1 Like Inhibitor	Animal Kazal-type Inhibitors	1pce 1hpt
			Virally Encoded KP4 Toxin	Virally Encoded KP4 Toxin	1kpt
	6 (0.51)	100% SP	Kringle Modules	Kringle Modules	1pk4 2pf1 1pml 1pkr
	51	SP	Kringle Modules	Kringle Modules	2hpqP
	52	SP	Insulin Like	Insulin Like	6ins
D2	53	SP	Kringle Modules	Kringle Modules	1kdu
	56	SP	Insulin Like	Insulin Like	1igl
	40	β	Thaumatin	Thaumatin	1thv
	22	SP	Fibronectin Type I Molule	Fibronectin Type I Molule	1fbr
	2 (0.48)	100% α	hospholipase A2	Insect Phospholipase A_2	1poc 1ppa 1poa 1pp2L
D3			Phospholipase A_2	Vertebrate Phospholipase A_2	1bp2 1pod
	54	SP	Protein Inhibitor PMC-C	Protein Inhibitor PMC-C	1pmc
D4	24	α	Hemocyanin	Hemocyanin	1hc4
E	1 (0.61)	100% All β	Trypsin Like Serin Protease	Eukariotic Proteases	1lmwB 3gctA 1ton 3rp2A 4ptp 3est 1ppfE 1bit 1hcgA
	17	β	Trypsin Like Serin Protease	Prokariotic Proteases	1arb

Continúa la Tabla IV.2

Cla	Cluster	Clase Estruct.	Tipo de Plegamiento	Familia	Código de PDB		
F1	4 (0.31)	100% α	Hemodependent Peroxidase	Citochrome C Peroxidase-Like	1lgaA	1arv	1mnp
	8 (0.51)	100% α+β	Lisozyme Like	Lisozyme Like	193l 1hml	2eql	1lzr
F2	27	α+β	Cysteine Proteinase	Papain Like	1hucB		
	48	α/β	α/β Hydrolase	Serin Carboxipeptidase	1cpy		
	31	α/β	α/β Hydrolases	Pancreatic Lipases	1lpbB		
	44	α/β	α/β Hydrolase	Fungal Lipase	3tgl		
	21	α/β	Flavodoxin Like	Esterase	1esc		
F3	36	α+β	Cysteine Proteinase	Papain Like	1ppn		
	57	α+β	Cysteine Proteinase	Papain Like	2act		
	42	β	8 Blader β–Propeller	Methanol D.H.	3aahA		
	34	β	Glycosil Asparaginase	Glycosil Asparaginase	1pgs		
	7 (0.65)	100% SP	BPTI Like	Small Kurnitz-type inch BPTI Like Toxin	1dtx 1aapA	1knt 1shp	1dtk
F4	49	SP	BPTI Like Anticoagulator FX_a	BPTI Like Anticoagulator FX_a	1tap		
	29	β	4 BLader β_Propeller	Hemopepsin Like Domain	1hxn		
	18	α	Bifunctional Inh. Lipid Transfer Protein 2S Album.	Bifunctional Protease α–Amilase Inh.	1bip		
F5	11 (0.91)	100% α+β	RNAsa Like	RNAsa Like	1onc	7rsa	1ang
	23	β	Lipocalins	Lipocalins	1hbq		

Continúa la Tabla IV.2

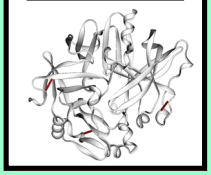
Cla	Cluster	Clase Estruct.	Tipo de Plegamiento	Familia	Código de PDB
		84.4% SP	Cysteine-Knot Cytokines	Neurotropin	1bet
		7.8% α		Gonadotropin	1hcnA 1hcnB
		4.7% β		Platelet Derived Grotwh Factor	1pdgA
		3.0% α/β		Transforming Growth Factor β	2tgi
			Epidermal Growth Factor Like Module	Epidermal Growth Factor Like Module	1prhA 1tpg 1hcgB 1zaq 4tgf 1hrf 3egf 1ccf
			Pheromone Protein	Pheromone Protein	1erp 1erl 1erd
			Small Inhibitors, Toxins and Lectins	Insect Antibacter. Protein	1ica
				Short chain Scorpion Toxin	1pnh 2crd 1mtx
				Large chain Scorpion Toxin	1nrb
				Plant γ–thionin	1gps
				(Pro)Colipase	1lpbA
				Spider Toxin	1oaw
				Plant Inhibitor Proteinase	4cpaI
				Agglutinin	1hev
G	3		Bifuctional Inhibitor Lipid Transfer Protein Storage 2-S Albumin	Plant lipid Transfer and hydrophobic protein	1hyp
	(2.57)		Thrombrin Inhibitor	Thrombrin Inhibitor	1dec 4hctI
			Bowman Birk Inhibitor	Bowman Birk Inhibitor	1pi2 2bbi
			Ovomucoid/PCI-1 Like Inhibitor	Plant Proteinase Inhibitor	4sgbI
			Snake Toxin Like	Snake Venon Toxin	1tgxA 1coe 1ntx 1fas 1kbaA 2abx 1ntn 1cvo 3ebx 1tfs
				Dendroaspin	1drs
			Defensin Like	Defensin	1atx 1dfnA 2sh1 1bnb 2bds 1ahl
			Anaphylotoxins	Anaphylotoxins	1c5a
			Crambrin Like	Crambrin Like	1cbn
			C-Type Lectin	C-Type Lectin	1esl
			Phospholipase A_2	Vertebrate Phospholipase A_2	1pp2L
			6 Bladed β–Propeller	Sialidases	1ncsA
			Methilamine DH	Methilamine DH	2madL
			β–Prims I	β–Prims I	1vmoA
			Periplasmatic Binding Protein	Transferrin	1lct
	37	β	C-Type Lectin	Pertussis Lectin S2/S3 Subunits	1prtB
	25	SP	Complement Control Module	Complement Control Module	1hfh

La Figura IV.2 muestra un resumen de la Tabla IV.2.

Figura IV.2 Clases obtenidas en el PDB tras utilizar la metodología detallada.

CLASE A

3pgs
Pepsin-like



Número de proteínas : 11
Fold principal: Acid Proteases

CLUSTER: PDB
10:1hrnA, 3psg, 3cms;
55:1htrB;**33:**1mhlC;**39:**1tca
45:6taa;**46:**1ack;**20:**1cnsA;
35:1ppi

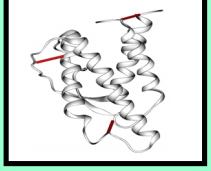
CLASE B

1tib
Fungal Lipase

Número de proteínas : 11
Fold principal: Periplasmatic Binding Protein like
CLUSTER: PDB
9:1lct,1nnt,1tfd;**50:**1tib,
30:1lki; **26:**1hgeA; **28:**1hvm;
16:1aozA;**58:**2fbjH; **41:**1vcaA

CLASE C

1rcb
Short Chain



Número de proteínas: 4
Fold principal:
Plant Cytotoxin

CLUSTER: PDB
15:1abr;**60:**2aaiB; **43:** 3gly
38:1rcb

CLASE D

1hylA
EukarioticProtease

Número de proteínas:
35
Fold principal:
D1: Trypsin Like Serine Proteases
D2 : Insulin-Like
D3 : Eukaryotic Proteases
D4 : Hemocyianin

CLUSTER: PDB
D1 **5:**1hylA, 1sgt, 3gctA,1try; **59:**2kaiB; **19:**1bw4;
47:1ate; **14:**2alp, 1ncsA,1poa; **32:**1lpt;
12:3sgbl, 1pce, 2bus;**13:** 1pce, 1hpt, 1kpt;

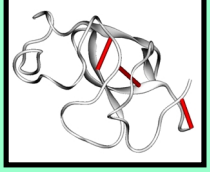
D2 **6:**1pk4, 2pf1, 1pml, 1pkr; **51:** 2hpqP; **52:**6ins;
53:1kdu; **56:**1igl; **40:**1thv; **22:** 1fbr;

D3 **2:**1poc, 1ppa, 1poa,1pp2L,1bp2,1pod;
54:1pmc;

D4 **24:** 1hc4;

CLASE E

1pk4
Kringle Module

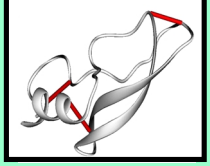


Number of Proteins : 10
Main Folds:·Trypsin Like Serine Proteases

CLUSTER: PDB
1: 1lmwB, 3gctA, 1ton, 3rp2A
4ptp, 3est, 1ppfE, 1bit,
1hcgA;
17:1arb;

CLASE F

1dtx
Small Kunitz-type



Número de proteínas : 27
Fold principal:
F1 : Citochromo C Peroxidase Like Family
F2 : Hydrolases.
F3 : Proteinases
F4 : Protein Inhibitors

CLUSTER: PDB
F1 **4:**1lgaA, 1arv, 1mnp;

F2 **8:**193l, 2eql, 1lzr, 1hml; **27:**1hucB; **48:**1cpy;
31:1lpbB; **44:**3tgl; **21:**1esc;

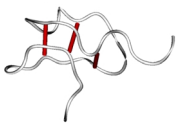
F3 **36:**1ppn; **57:**2act; **42:**3aahA; **34:**1pgs;

F4 **7:** 1dtx, 1knt, 1dtk, 1aapA, 1shp; **49:**1tap;
29:1hxn; **18:** 1bip;

F5 **11:**1onc, 7rsa, 1ang; **23:**1hbq;

CLASE G

4cpal
Plant Inhibitor



Número de protéinas: 66
Fold principal:
· Cysteine-Knot Cytokines
· Epidermal Growth Factor-like.
· Small Inhibitor, Toxins and Lectins
· Snake Toxin Like.
· Defensin Like.

CLUSTER: PDB
3:1bet, 1hcnA, 1hcnB, 1pdgA, 2tgi, 1prhA,1tpg,
1hcgB, 1zaq, 4tgf, 1hrf, 3egf, 1ccf, 1erp,1erl,
1erd, 1ica, 1pnh, 2crd, 1mtx, 1nrb, 1gps, 1lpbA,
1oaw, 4cpal, 1hev, 1hyp, 1dec, 4hctl, 1pi2, 2bbi,
4sgbl, 1tgxA, 1coe, 1ntx, 1fas, 1kbaA, 2abx,
1ntn, 1cvo, 3ebx, 1tfs, 1drs, 1atx, 1dfnA, 2sh1,
1bnb, 2bds, 1ahl, 1c5a, 1cbn, 1esl, 1pp2L,
1ncsA, 2madL, 1vmoA, 1lct;
37:1prtB; **25:**1hfh;

Procedemos a la descripción de cada una de las clases:

CLASE A: Esta clase se compone de 8 *clusters* y 11 proteínas. El 40% de ellas muestran el plegamiento de las proteasas ácidas y otro 30% son α/β hidrolasas.

CLASE B: Está también formada por 8 *clusters* y 11 proteínas. Las proteínas agrupadas son en un 50% α/β mientras que el 40% son todo-β. El principal *cluster* de la clase es el cluster 9, compuesto por 3 proteínas transportadoras de hierro.

CLASE C: Esta clase está formada por 4 proteínas cuyo plegamiento más representativo es el de β-*trefoil*.

CLASE D: Dieciocho *clusters* y 35 proteínas forman esta clase. Esta clase ha sido dividida en 3 subclases D1, D2 y D3.

SUBCLASE D1: Esta subclase muestra una distribución dicotómica. Una rama está formada por más del 80% de proteínas todo-β. El 70% de estas proteínas tienen actividad catalítica, siendo el principal tipo de plegamiento el *Trypsin-like Serine Proteases*. El 100% de la otra rama son *Small Proteins* compuestas por dos *clusters* (7 proteínas) con función inhibidora.

SUBCLASE D2: Está compuesta por 10 proteínas, donde cabe destacar dos plegamientos distintos: *Insuline-like y Kringle Modules*, los cuales incluyen el 90% de estas proteínas.

SUBCLASE D3: La mayoría de las proteínas de esta subclase muestran un plegamiento todo-α (más del 90% de las proteínas), siendo un 75% de la familia *Phospholipase* A2.

CLASE E: En esta clase sólo un plegamiento ha sido detectado, el correspondiente a *Eukaryotic Proteases*. La mayoría de las proteínas son proteasas de vertebrados. El principal *cluster* es el número 1, compuesto por plegamiento todo-β.

CLASE F: Está formada por 17 clusters y 27 proteínas distribuida de un modo politómico.

SUBCLASE F1: Está formada por 3 proteínas con el mismo plegamiento *(Hemoperoxidases)* pertenecientes a la *cytochrom C Peroxidase-like Family*.

SUBCLASE F2: Está principalmente formada por hidrolasas. El 80% presentan α/β como plegamiento. En esta subclase, hay dos tipos de familias de proteínas: *α/β hydrolases y Lysozime-like*. El primero es funcionalmente más heterogéneo y sus proteínas son especialmente carboxipeptidasas, esterasas y lipasas. Sin embargo, esta divergencia funcional no se aprecia en la topología de puentes de azufre. El principal *cluster* de esta subclase es el número 8 formado por lisozimas y la lactalbúmina.

SUBCLASE F3: Esta constituida por *Cysteine Proteinases* y miembros de la familia *Papain-like*. Algunas de estas proteínas están también presentes en la subclase F2 pero con diferente disposición de puentes de azufre.

SUBCLASE F4: Está compuesta por inhibidores enzimáticos y *Small Proteins* (75%). El principal *cluster* es el número 7, el cual está formado por 5 proteínas.

CLASE G: Esta clase contiene el mayor número de proteínas. El principal *cluster* (número 3) se une a dos clusters más a bastante altura en el dendrograma, para acabar de definir esta clase. Es importante enfatizar que en el *cluster* 3 hay proteínas con muy diferente función y estructura, pero que el 85% de ellas son *Small Proteins* contenidas en 17 plegamientos diferentes. Cinco de estos 17 plegamientos (*Cysteine-Knot citokines, EGF-like, Small inhibitor toxins and lectins, Snake toxin Like y Defensin-like*) representan más del 50% del total de proteínas del *cluster*. La alta equivalencia topológica de los puentes de azufre y el elevado número de proteínas contenidas en el *cluster* 3 hace de él el *cluster* más interesante para la realización de determinados estudios. De hecho, este cluster contiene un importante número de proteínas

con la conocida topología β-*cross,* la cual ha sido estudiada por otros autores (Harrison et al. 1994 y 1996).

Comparación entre proteínas de la misma clase (dentro de clase)

La clasificación por topología de puentes de azufre divide el sistema inicial de estudio en la clasificación estructural obtenida (Tabla IV.2 y Tabla IV.3). No es sorprendente que proteínas homólogas de la misma familia o similar plegamiento queden unidas en esta clasificación. Sin embargo, es destacable que algunas de estas familias de proteínas puedan separarse en clases distintas. También es posible observar que proteínas con más de 3 puentes de azufre pueden aparecer en más de una clase, utilizando para ello topologías de puentes de azufre distintas. Varios casos destacables han sido analizados cuidadosamente, detallándose los resultados obtenidos a continuación:

- El *cluster* 5 (en la subclase D1) y el *cluster* 1 (en la clase E) están en clases distintas. Sin embargo, todas estas proteínas son de la familia *Serine Proteases*. Cuando analizamos cada proteína, podemos ver que el *cluster* 5 está compuesto por proteínas de vertebrados mientras las proteínas formadas por el *cluster* 1 son de hongos o gusanos, es decir, alejados evolutivamente. No obstante, la proteína 3gct *(γ-Chymotrypsin* de pancreas bovino) está presente en los dos clusters utilizando una diferente combinación de puentes de azufre.

- Las subclases F2 y F3 presentan proteínas con el mismo tipo de plegamiento: *Papain-like*. Usando nuestra metodología se han agrupado en distintas clases. Analizando las proteínas incluidas en estas clases vemos que ambas quedan muy separadas evolutivamente. La subclase F3 está compuesta por proteínas de kiwi o papaya mientras las proteínas de la subclase F2 son proteínas humanas.

Comparación de proteínas agrupadas en el mismo cluster (dentro de cluster)

El *cluster* más destacable es el *cluster* 3, enmarcado en la clase G. Está formado por 64 proteínas con alta equivalencia topológica de sus puentes de azufre (ssRMSD=2.6 Å). La mayoría de ellas son *Small Proteins*, con poca o ninguna estructura secundaria regular, y mayoritariamente con 3 o 4 puentes disulfuro. Una gran cantidad de proteínas incluidas en este *cluster* son factores de crecimiento, hormonas inhibidores y venenos. Aparecen también proteínas con elevado número de puentes de azufre y con gran cantidad de estructuras secundarias regulares, como la 1pp2L (una fosfolipasa), las cuales hacen aumentar la heterogeneidad de la clase.

Han sido estudiadas las relaciones estructurales y funcionales entre proteínas de distintas familias de este *cluster*. Estos estudios se han realizado mediante la superposición de las proteínas por topologías de puentes de azufre. Se ha observado que, incluso entre proteínas de distintas familias, aparecen similitudes estructurales a todas ellas, como son: propiedades fisicoquímicas de las cadenas laterales, la estructura de la cadena principal y estructuras secundarias regulares. Estos efectos han sido ya descritos por otros autores definiéndose algunas de estas características como importantes desde un punto de vista estructural y funcional (Groenen et al., 1994, Jacobsen et al., 1996; Barbacci et al., 1995; Nogata et al., 1994; Picot et al., 1994; Montelione et al., 1992; Uller et al., 1992; Chang et al., 1994 y 1995; Brown et al., 1989 y 1992).

Las proteínas que forman el *cluster* 3 han sido analizadas bajo condiciones aún más restrictivas por el KNOT-MATCH. Los resultados obtenidos mantienen las mismas características que los encontrados para el análisis de todo el sistema. Es decir, proteínas de la misma familia tienden a unirse, observándose características estructurales comunes. Un análisis secuencial y conformacional de estas proteínas muestra cómo elementos de la misma familia son, en algunos casos, más parecidos comparando la topología de sus puentes de azufre que si se realiza un simple alineamiento secuencial. (Resultados no presentados).

Comparación de proteínas de la misma familia (dentro de familia)

En este estudio hemos incluido 2 familias: *EGF-like* y *Defensin-like*. La comparación entre proteínas de este *cluster* ha sido realizada por comparación de cada miembro de la familia frente al resto. El análisis ha implicado la detección de propiedades fisicoquímicas de las cadenas laterales conservadas en el espacio para la mayoría de las proteínas de la familia. Este análisis se ha realizado previa superposición de las proteínas por la topología de los puentes disulfuro. Las propiedades analizadas son la conservación secuencial/espacial de residuos aromáticos, cargados, hidrofóbicos, polares o voluminosos. Este estudio se ha realizado para los 7 miembros de la familia *EGF-like* y para los 6 elementos de la familia *Defensin-like* por separado.

En la familia *EGF-like* han sido encontradas 15 posiciones con propiedades fisicoquímicas equivalentes (representadas en más del 57% de las proteínas de la familia) (ver Figura IV.3). Nueve de las 15 posiciones han sido descritas por otros autores como funcionalmente importantes. Estos residuos, según la enumeración del EGF son: Y13, N16, I23, S25, Y29, N32, Y37, R41 y R45. La Tabla IV.3 contiene un resumen de las funciones de estos residuos.

Residuo	*Descripción funcional*
Y13	Importante en la unión receptor EGF. Implicado en la respuesta mitogénica. En la heregulina el F18 (el residuo equivalente) es responsable de la formación de un bolsillo hidrofóbico implicado en la unión a receptor.
N16	Este residuo y el equivalente en la heregulina (N192) puede tener importancia en la unión con el receptor.
I23	Importante en la interacción con el receptor.
S25	Importancia estructural.
N32	Descrita como bisagra en algunas moléculas.
Y37	Descrito en el EGF como muy importante para la unión a receptor.
Y29	Descrita la Y197 de la heregulina (residuo equivalente) como esencial para la formación de un bolsillo hidrofóbico necesario para la unión al receptor.
R41	En la heregulina la R220 (posición equivalente) se describe como esencial para el reconocimiento del receptor.
R45	En el TGF juega un importante papel en la unión al receptor.

Tabla IV.3: Análisis de residuos que conservan las propiedades fisicoquímicas en el espacio al superponer todas las proteínas de la EGF-Like Family por la topología de sus puentes de azufre.

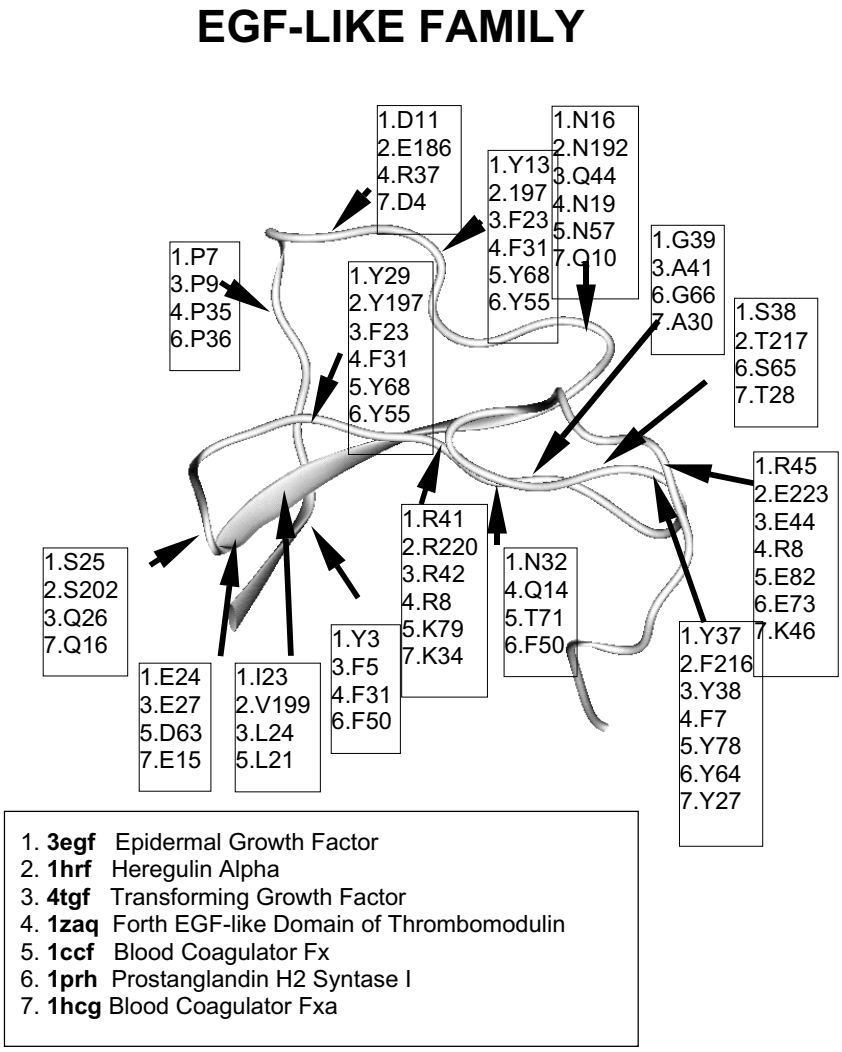


Figura IV.3: Esqueleto polipeptídico del EGF. Este esqueleto es comparable estructuralmente al resto de las proteínas de la misma familia. La flechas marcan la posición de los residuos que tienen propiedades fisicoquímicas comunes en el espacio y son importantes funcionalmente.

La familia de las *Defensin-like* presenta igualmente varias posiciones donde las cadenas laterales de los residuos de las distintas proteínas confieren al espacio propiedades fisicoquímicas equivalentes. Sin embargo, al ser una familia de proteínas no tan estudiadas como la *EGF-like* family, la construcción de una tabla equivalente a la Tabla IV.3 ha resultado imposible. A pesar de ello, la Figura IV.4 muestra esquemáticamente aquellos residuos que se conservan en el espacio.

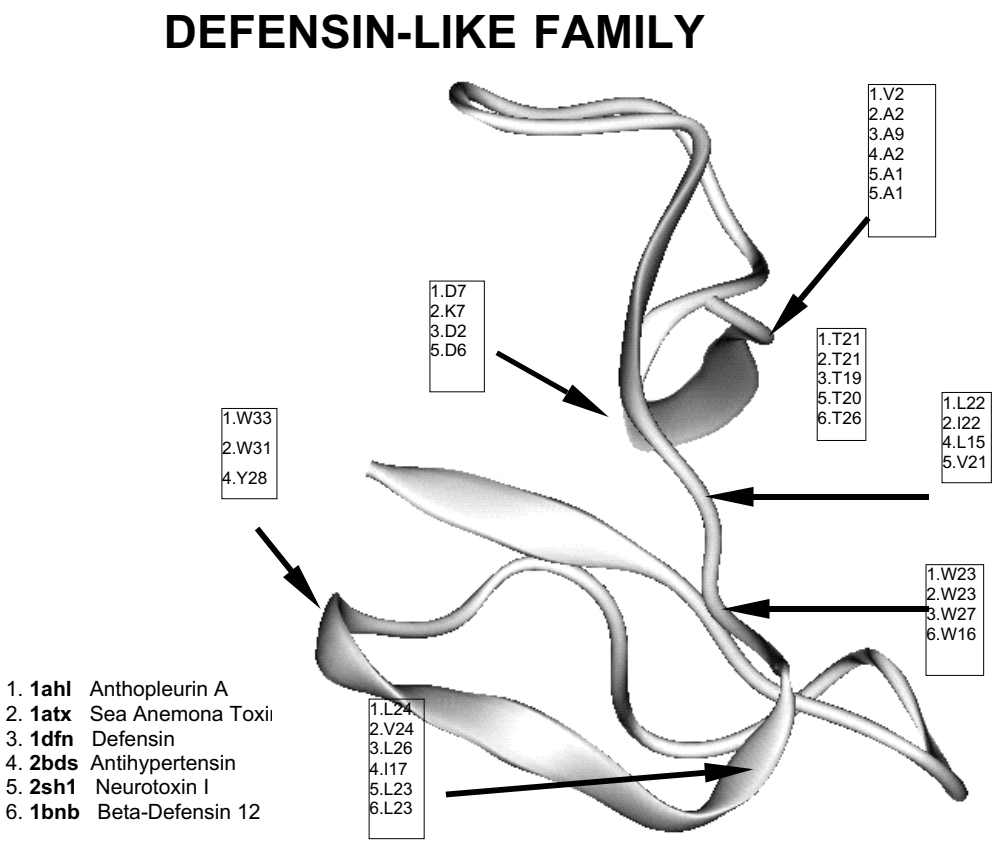


Figura IV.4 Esqueleto polipeptídico de la defensina. Se marcan las posiciones del espacio donde las proteínas de esta familia conservan las propiedades fisicoquímicas.

Comparación de proteínas de distinta familia en el mismo cluster (dentro de cluster, entre familias)

Se han realizado determinados estudios a fin de observar relaciones estructurales entre proteínas del mismo cluster pero que no forman parte de la misma familia. Como ya ha sido comentado, las familias, según las clasificaciones SCOP y CATH (Murzing et al., 1995; Orengo et al., 1997), se establecen teniendo en cuenta sobretodo la información secuencial y las posiciones relativas y tipos de estructuras secundarias regulares de las proteínas. Dado que se está realizando una clasificación puramente estructural, donde no se tienen en cuenta estos factores, cabe esperar que aparezcan proteínas que tengan elementos estructurales, (lazos, regiones, etc.) comunes, y que no presenten una equivalencia secuencial, quedando por tanto estos resultados ocultos a los algoritmos de alineamiento clásicos.

La observación de relaciones estructurales entre proteínas que secuencial, funcional y evolutivamente no necesariamente tienen que estar relacionadas ha sido descrita anteriormente (Madej et al., 1995; Mas et al., 1998). Tal estudio merece por tanto un análisis profundo que será abordado en el siguiente capítulo de este trabajo.

IV-A.5 Discusión y conclusiones

Hemos utilizado la topología de puentes de azufre para facilitar la superposición estructural de proteínas ricas en puentes disulfuro sobre un PDB no redundante. Hemos creado varios algoritmos y hemos construido un programa, KNOT-MATCH, para definir las distintas topologías de puentes que muestra el sistema de estudio. Se ha ejecutado el programa en las condiciones pertinentes y se han analizado los resultados. El análisis de los resultados muestra ciertas regularidades en la distribución topológica de los puentes de azufre. Hemos estudiado si estas preferencias por determinadas topologías pueden ser utilizadas para agrupar proteínas y estudiar las posibles relaciones entre ellas. Así, hemos

construido una clasificación preliminar de proteínas ricas en puentes de azufre y hemos dividido el conjunto inicial de proteínas del que partíamos en 7 clases. El criterio para la separación de las proteínas en clases era esencialmente la topología de los puentes de azufre. Estas clases contienen proteínas de distintas familias, aunque es muy común encontrar todos los miembros de una familia en una clase, como por ejemplo la familia de las fosfolipasas A2 en la Clase A o la *EGF-like Family* en la clase G.

La topología de puentes de azufre

A la vista de los resultados queda claro que es posible hacer agrupaciones de proteínas basadas en las topologías de sus puentes de azufre. Este hecho indica que existen maneras preferenciales de disponer los puentes de azufre en el espacio. En algún caso ha sido posible establecer relaciones evolutivas entre proteínas, pese a no ser proteínas con alta homología secuencial. No obstante, aparecen casos donde la relación evolutiva es mucho más confusa, invitando a pensar en una evolución convergente (Mas et al., 1998). Este fenómeno nos hace pensar que la disposición de los puentes de azufre en una determinada forma tiene una explicación fisico-química y energética que hace que se dispongan de formas preferenciales en el espacio. Este hecho ya fue comentado por Harrison & Sternberg en 1996 con referencia a unas estructuras, los β-*cross*, en las que aparecían dos puentes de azufre como elementos estructurales de gran importancia.

Las clasificaciones generales

Como ya ha sido comentado, las clasificaciones generales de proteínas se basan esencialmente en la información secuencial y en la estructura secundaria de las proteínas. Sin embargo, estas clasificaciones tienen un gran interés para poder establecer relaciones entre proteínas. Tal hecho facilita la adaptación de proteínas para aplicaciones heterólogas, o bien la de sus dominios o elementos estructurales de función conocida. Las proteínas con poca o ninguna estructura secundaria regular son obviadas o tratadas de un modo especial en las

clasificaciones tradicionales de proteínas, formando por tanto un grupo de proteínas difícilmente relacionables entre ellas y/o con ellas. Hemos visto incluso cómo algunas de estas proteínas en el momento de la realización de este trabajo no son catalogadas por CATH, como son: 1bnb, 1dec, 1dfnA, 1erp, 1mtx, 1pnh, 1prhA, 2crd, entre otras. La información obtenida por SCOP (Murzin et al, 1995) o desde el PDB (Bernstein et al, 1977) nos ha permitido relacionarlas, pudiendo observarse que estas proteínas fueron agrupadas por nuestra metodología con miembros de su familia funcional y/o con proteínas con grandes similitudes estructurales. En otros casos, CATH y SCOP muestran una clasificación claramente diferente para este tipo de proteínas, nótese el caso del PCI clasificado como *Small Protein* o como todo β por CATH y SCOP respectivamente. Es por tanto de gran interés disponer de herramientas como el KNOT-MATCH, que permite estudiar este tipo de proteínas desde un punto de vista puramente estructural.

Redefinición de Familias de Proteínas.

Un ejemplo de proteínas ricas en puentes de azufre son aquéllas que cumplen el llamado motivo *T-Knot* (Sun, 1995; Isaacs, 1995). Varios autores han descrito el *T-Knot* a partir de la unión secuencial de cisteínas, es decir, aquellas proteínas que muestran la formación de sus puentes de azufre según el patrón de unión de sus cisteínas (1-3, 2-4 y 3-6) (Lin et al., 1995); dicho de otra manera, proteínas cuya primera cisteína forma puente disulfuro con la tercera, segunda con la cuarta y tercera con la sexta. Otros autores (Harrison & Sternberg, 1996) han agrupado algunas de estas proteínas a partir de características estructurales como son dos puentes de azufre y una hoja β, llamando a esta estructura *β-cross*.

Nuestra aproximación se basa exclusivamente en la topología de 3 puentes de azufre, donde no se tienen en cuenta ni el orden secuencial de unión de cisteínas ni la presencia/ausencia de estructuras secundarias regulares. A pesar de estas diferencias aparecen una gran cantidad de proteínas comunes entre nuestro criterio y los dos anteriores. Por ejemplo, el PCI y el EGF. Estas proteínas han sido agrupadas juntas por nuestra metodología, sin embargo no tienen un orden

secuencial común de unión de cisteínas, ya que el PCI tiene uniones 1-4, 2-5, 3-6 y el EGF presenta uniones 1-3, 2-4, 5-6; por otro lado, el EGF tiene una hoja β formada por 5 residuos en cada *strand* o cadena, pero el PCI tiene sólo 2 residuos potencialmente implicados en la formación de una banda de una hoja β. Muchos autores consideran que ésta es una longitud insuficiente para ser considerada como estructura secundaria regular (Kabash & Sander, 1983). Por otra banda, ambas proteínas tienen similares regiones formando lazos (7 residuos con menos de 0.8 Å de RMSD) y presentan algunas propiedades físico-químicas altamente conservadas en el espacio. Nuestra definición de T-Knot puede ser complementaria a las aproximaciones hechas por otros autores, y en algunos casos, puede ser más general.

Implicaciones en Plegamiento y Estructura de este tipo de Proteínas.

Para proteínas pequeñas, como el PCI, la importancia de los puentes de azufre en el plegamiento y estabilidad de su estructura ha sido demostrada por diversos autores (Chatrenet, 1992; Chang, 1994 y 1995; Isaacs, 1995). En nuestra clasificación podemos ver cómo proteínas con y sin homología secuencial, pero incluidas en la misma clase, presentan mayoritariamente características estructurales comunes. Así, relaciones estructurales no evidentes pueden ser extraídas a partir de alineamientos de la topología de puentes de azufre, y pueden ayudar a entender determinados tipos de plegamientos de proteínas. Algunos autores (Harrison & Sternberg, 1996) definen la importancia de los *β-cross* en el plegamiento de estas proteínas, estructura que incluye dos puentes de azufre.

Los resultados obtenidos muestran también que proteínas con conformaciones similares pueden no tener homología secuencial y presentar similitudes en la estructura de la cadena peptídica, tras superponer estas conformaciones por la topología de sus puentes de azufre. Así, la topología de puentes de azufre puede ser útil para la comparación estructural de proteínas ricas en tales enlaces, permitiendo, en muchos casos, obtener un alineamiento correcto de sus estructuras secundarias regulares (o de una parte significativa de ellas), incluso cuando se aplica entre miembros de diferentes familias. La aproximación que

proponemos aquí puede ser, por tanto, de utilidad general en el estudio de relaciones estructura/función de proteínas ricas en puentes de azufre.

IV-B RELACIONES ESTRUCTURALES EN PROTEÍNAS HALLADAS A PARTIR DE LA SUPERPOSICIÓN DE LAS TOPOLOGÍAS DE SUS PUENTES DE AZUFRE.

La superposición de proteínas es un importante procedimiento para establecer relaciones estructurales y funcionales. El programa KNOT-MATCH, cuyo funcionamiento ha sido descrito en los apartados anteriores, ha sido preparado para obtener superposiciones estructurales de proteínas de un modo automático a partir de las topologías de sus puentes de azufre. Como resultado de estas superposiciones las estructuras secundarias regulares, los *loops* (lazos de proteínas) y las cadenas laterales de determinados residuos son alineados correctamente. Estos alineamientos nos han permitido superponer en el espacio residuos con similares características físico-químicas, algunos de ellos descritos en la literatura como funcionalmente importantes para una o más proteínas de una familia, y también entre proteínas aparentemente no relacionadas, pero agrupadas juntas utilizando nuestra metodología (ver el apartado anterior).

Familias de proteínas ricas en puentes de azufre, como la familia *EGF-like,* la familia *defensin-like* y el PCI, único miembro tratado en esta tesis de la familia *Plant inhibitor proteinase*, han sido analizados profundamente a partir de sus superposiciones. Se ha observado que, tras superponer dos proteínas de estas familias mediante el KNOT-MATCH, algunos aminoácidos presentan en la superficie de la proteína grupos radicales con similares propiedades fisicoquímicas. El programa nos ha permitido relacionar estas proteínas de tal manera que por técnicas clásicas de alineamientos de proteínas no sería posible vincular.

IV-B.1 Introducción

Comparación de proteínas

Los algoritmos y programas para el alineamiento secuencial de proteínas son frecuentemente utilizados como herramientas en biotecnología (para más información consultar la revisión realizada por Adrade & Sander en 1997). Un marco similar acontece hoy para el alineamiento tridimensional (3D) y análisis de proteínas. El ritmo actual al cual nuevas estructuras de proteínas son conocidas crece día a día gracias a los avances en las técnicas de cristalografía de Rayos X y a la Resonancia Magnética Nuclear. Estos avances demandan nuevos y más rápidos algoritmos para la comparación 3D de proteínas. Dado que las estructuras 3D son muy conservadas en la evolución (Rost, 1997), su comparación de estructuras 3D permite establecer relaciones entre proteínas que parecen no estar relacionadas o que quedan ocultas tras aplicar alineamientos secuenciales. Los primeros métodos diseñados para este propósito, el alineamiento 3D de proteínas, requerían de un alineamiento secuencial previo, eran muy lentos y estaban limitados a proteínas homologas (Rossman & Argos, 1976; Mathews & Rossman, 1985). En los últimos años se ha desarrollado una nueva generación de algoritmos. Los más eficientes de ellos permiten un alineamiento automático y rápido facilitando el rastreo de las bases de datos (Brenner, 1995). La superposición estructural de proteínas requiere el apareamiento de residuos o regiones equivalentes[1] de las proteínas para regir la superposición de sus estructuras. El número de formas en las cuales se puede superponer dos proteínas es casi infinito, y frecuentemente se utilizan los alineamientos secuenciales o la coincidencia de estructuras secundarias regulares para guiar tal superposición. Sin embargo, cuando las proteínas carecen de estructuras secundarias regulares o bien sus dimensiones son insuficientes, y no hay homología secuencial entre ellas, el alineamiento estructural resulta muy complejo, y es necesario el desarrollo de nuevas técnicas que permitan obtener alineamientos correctos en tales casos (Johnson et al., 1994; Srinivasan et al., 1996).

[1] en el sentido utilizado en este trabajo para definir la equivalencia entre átomos

Proteínas de interés biomédico y biotecnológico

Un gran número de proteínas con importantes funciones biológicas (factores de crecimiento, hormonas, toxinas, etc...) son proteínas que contienen una importante cantidad de puentes de azufre que determinan su plegamiento y tendencias topológicas (Creighton, 1992; Chang et al., 1994 y 1995; Wu et al., 1998). El correcto plegamiento de la topología de sus puentes de azufre es indispensable para que la proteína adopte su estructura nativa y por tanto pueda desarrollar su función (Betz, 1993). En este contexto, Thornton en 1980 y Richardson en 1981 utilizaron el orden secuencial de las cisteínas para establecer una clasificación preliminar de proteínas ricas en puentes de azufre. Más recientemente se han establecido otras clasificaciones que tenían en cuenta aspectos estructurales, como los β-*cross* (una hoja β sostenida dos puentes de azufre) utilizados por Harrison y Sternberg en 1996 para establecer una nueva clasificación. Estas proteínas ricas en puentes de azufre suelen ser pobres en estructuras secundarias regulares y su alineamiento en el espacio podría ser muy útil para poder establecer relaciones estructurales y funcionales entre ellas. El KNOT-MATCH se ha mostrado como una herramienta que ofrece grandes posibilidades y perspectivas para este tipo de análisis.

IV-B.2 Objetivos

El principal objetivo de este apartado es determinar similitudes estructurales y justificar y predecir similitudes funcionales, entre proteínas. Este objetivo se desarrollará en primer lugar para proteínas de una misma familia, y finalmente, entre proteínas no relacionadas por su secuencia, ni por su estructura ni por su función. Todos los análisis propuestos utilizarán la superposición de las estructuras obtenidas tomando como elemento estructural común la topología de sus puentes de azufre.

IV-B.3 Métodos

Para evaluar la importancia estructural de la topología de puentes de azufre y poder establecer otras relaciones estructurales a partir de la arquitectura de los puentes de azufre, se han superpuesto proteínas de distintas familias y orígenes evolutivos. La Figura IV.5 muestra tal superposición. En algunos casos, se han podido establecer relaciones estructurales entre las cadenas laterales de determinados residuos y/o el esqueleto polipeptídico de las proteínas. El conjunto de coincidencias en el espacio 3D ha sido contrastado con datos experimentales publicados en la literatura. La superposición de las proteínas permite observar la preferencia de estas proteínas por una determinada disposición de puentes de azufre. Este hecho ha sido utilizado para agrupar las proteínas del PDB, estableciéndose así la clasificación preliminar de proteínas ricas en puentes de azufre presentada anteriormente. En algunos casos, las superposiciones obtenidas por el KNOT-MATCH permiten un correcto alineamiento de las estructuras secundarias regulares y/o la disposición espacial de determinados residuos, incluso entre proteínas aparentemente no relacionadas.

IV-B.4 Resultados

Análisis de proteínas de la misma familia

La familia de los *EGF-like*, definida por la clasificación general de proteínas SCOP (Murzing et al., 1995), presenta un nudo de cisteínas que ha sido descrito como estructuralmente importante, e incluso como elementos que determinan el proceso de plegamiento de estas proteínas (Wu et al.,1998). De algunas de las proteínas de la familia, dada su importancia biológica, se dispone de abundante información referente al papel estructural y funcional de algunos residuos. Este hecho y la gran variabilidad estructural entre los miembros de esta familia hacen de ella un excelente modelo para profundizar en el análisis entre miembros de una misma familia (**dentro de familia**), a partir de la superposición obtenida con nuestra metodología. Tras superponer las proteínas con el KNOT-MATCH, todos los miembros de la familia fueron comparados por parejas para extraer,

mediante un exhaustivo análisis con el programa TURBO-FRODO, un *consensus* de las propiedades fisicoquímicas que las cadenas laterales disponían en el espacio (Rousell et al., 1994). Se determinaron 15 posiciones de aminoácidos cuyas propiedades fisicoquímicas coincidían en más del 55% de las proteínas de la familia (Figura IV.3). Como se esperaba, tras la superposición de miembros de la misma familia, el *consensus* obtenido mostraba sustituciones secuencial y espacialmente conservativas. Algunos de los residuos detectados por la metodología utilizada aparecían descritos en la literatura como estructural y/o funcionalmente importantes, descritos por diversos autores, tras utilizar estudios de mutagénesis dirigida y otras aproximaciones (Groenen et al., 1994; Barbacci et al., 1995). Sin embargo, aparecían otros residuos, como la Prolina 7 del EGF, que no había sido previamente descrita ni como funcional, ni como estructuralmente importante para esta familia de proteínas y que sin embargo se mostraba común para un importante número de estas proteínas (Tabla IV.3). El *consensus* obtenido podía por tanto desvelar nuevos residuos estructural y/o funcionalmente importantes. Resultados similares se obtenían cuando este procedimiento se repetía para la familia de las defensinas (Figura IV.4).

Las defensinas son proteínas con función antimicrobiana y en cuya estructura proteica cabe destacar la presencia de tres puentes de azufre y una clara hoja β. Existen tres categorías de defensinas, las de invertebrados (principalmente insectos) y los tipos α y β de vertebrados (White et al., 1995; Gantz & Leher, 1998). Los tres grupos difieren de la conectividad de sus seis cisteínas pero no en la topología de los puentes que éstas forman, tal y como se ha podido ver con el KNOT-MATCH. Tras la superposición de seis miembros de esta familia: antopleurina-A (1ahl), toxina de anémona marina (1atx), defensina-HNP3 (1dfnA), antihipertensina (2bds), neurotoxina I (2sh1) y β-defensina (1bnb), siete residuos conservaban la mismas características fisicoquímicas en el espacio mediante la superposición realizada con el KNOT-MATCH. Además se observaba un correcto alineamiento de elementos de estructura secundaria regular.

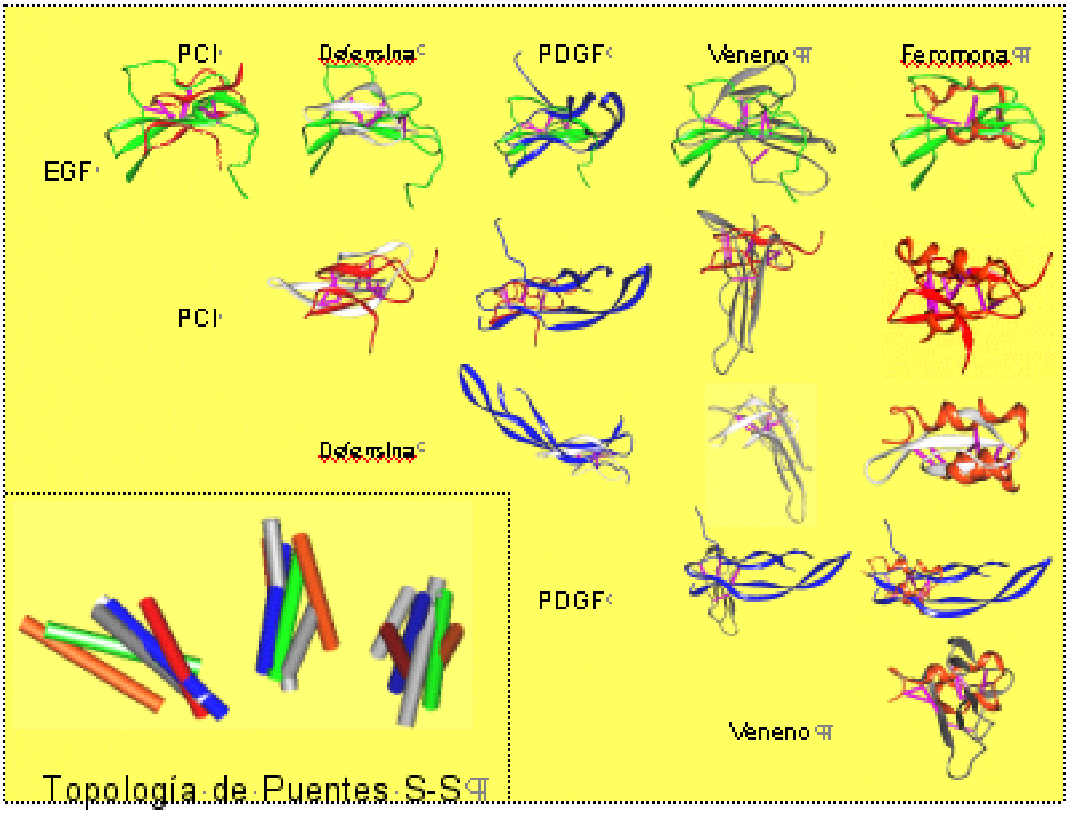


Figura IV.5 Se muestra la superposición de proteínas con baja homología secuencial por las topologías de sus puentes de azufre. Se puede apreciar la distribución de los puentes de azufre de estas proteínas. El promedio de RMSD es de 2.37 Å, lo cual indica que la superposición de estas estructuras es bastante buena, es decir, que la arquitectura de los puentes de azufre de estas proteínas son muy similares. Las funciones y orígenes evolutivos de las proteínas son muy diversos, siendo hormonas, factores de crecimiento, venenos e inhibidores de proteasas. Las abreviaciones corresponden a EGF (factor de crecimiento epidérmico), PDGF (factor de crecimiento y desarrollo plaquetario) y PCI (inhibidor de carboxipeptidasa de patata).

Análisis de proteínas aparentemente no relacionadas

La superposición de proteínas por la topología de sus puentes de azufre nos ha permitido encontrar similitudes, no sólo entre proteínas claramente relacionadas (por ejemplo lo observado entre proteínas de la misma familia), sino también entre proteínas que aparentemente no estaban relacionadas ni por sus secuencias ni por sus funciones. Este fenómeno es de gran interés, ya que nos permite relacionar estructuralmente dos proteínas de funciones no necesariamente similares, y por tanto, permite inferir sobre la función heteróloga de estas proteínas.

Hemos estudiado en detalle el caso de dos proteínas que no se encontraban en esta situación, no habían sido relacionadas previamente ni desde un punto de vista secuencial ni desde un punto de vista funcional. Este es el caso del PCI y del EGF. El último capítulo de este trabajo de tesis trata detalladamente la información estructural y funcional obtenida tras la comparación de estas dos proteínas, por lo tanto trataremos aquí estas proteínas como elementos estructurales, sin tratar las claves funcionales que las unen.

El PCI es un inhibidor de carboxipeptidasa de origen vegetal cuya función biológica podría estar relacionada con los mecanismos de defensa frente a insectos. El EGF es una proteína de origen animal que actúa como factor de crecimiento estimulando la división celular en tejidos epiteliales. Estas dos proteínas son claramente diferentes desde un punto de vista funcional y evolutivo pero también desde un punto de vista secuencial y estructural. El PCI tiene 39 residuos frente a los 53 del EGF. El EGF consta de una hoja β común a todos los miembros de su familia mientras el PCI contiene tan sólo 5 residuos en una disposición y puentes de hidrógeno similar a la que los aminoácidos adoptan en tal estructura secundaria, siendo este número de residuos insuficiente para muchos autores para ser considerada una hoja β (Ripscon et al., 1972). Estas proteínas tienen en común el hecho de disponer de tres puentes de azufre, siendo diferente el patrón secuencial de unión de las cisteínas para formar los puentes de azufre. Así el patrón de unión del PCI es 1-4, 2-5, 3-6 (la primera cisteína de la secuencia se une a la cuarta y así sucesivamente) mientras el

patrón que describe el EGF es 1-3, 2-4, 5-6. Estas diferencias estructurales y secuenciales hacen muy difícil la superposición de sus estructuras mediante el uso de métodos clásicos para permitir un análisis más profundo. El KNOT-MATCH aprovecha el único elemento claramente común entre ellas, la topología de sus puentes de azufre para superponer sus estructuras.

El análisis de estructural de estas proteínas tras su superposición por el KNOT-MATCH ofrece patrones estructurales comunes entre estas proteínas hasta ahora no descritos, como son:

1. correcto alineamiento de la topología de los puentes de azufre, determinado previamente por el KNOT-MATCH en la elaboración de la clasificación de proteínas ricas en puentes de azufre

2. las hojas β de ambas proteínas, si consideramos la existencia de esta estructura secundaria en el PCI, quedan alineadas en el espacio

3. dos lazos destacables en el PCI quedan alineados con dos lazos de similares características estructurales en el EGF

4. las cadenas laterales de 10 residuos superpuestas en el espacio muestran en la superficie de ambas moléculas similares propiedades fisicoquímicas, habiendo sido algunas de estas posiciones descritas por otros autores como funcional o estructuralmente importantes (Rees & Lipscomb, 1982; Brown et al., 1989; Brown & Wüthrich, 1992; Ullner et al., 1992; Groenen et al., 1994; Picot et al., 1994; Molina et al., 1994; Barbacci et al., 1995; Jacobsen et al., 1996; Blanco-Aparicio et al., 1998).

Las coincidencias estructurales de estas proteínas nos invitó a analizar más profundamente otras familias de proteínas que habían sido relacionadas mediante el KNOT-MATCH. De este modo se determinó que la región C27-C34 del PCI se superponía con bajo RMSD con la región G26-H34 de la angiotoxina de escorpión y con el lazo K27-R34 de la caribdotoxina (aproximadamente 0,7Å). Por su parte, la región H22-Y29 del factor de coagulación X se superponía a esta misma región con un RMSD de 1,8 Å mientras otro miembro de su familia de

proteínas, el EGF, lo hacía con un valor de 0,7 Å al alinearse con la región H22-Y29. Otros lazos del PCI fueron también estudiados con cierto detalle, así se determinó que el lazo C18-C24 era superponible con el lazo K47-C54 del veneno de serpiente (1coe) con un RMSD de 0,6Å. Similares resultados se obtenían entre los miembros de la familia de los *EGF-like*, así, el factor de coagulación X y el EGF, ambos con el lazo K47-C54, mostraban RMSD próximos a los 0,5 Å . Cabe destacar que los métodos clásicos de alineamiento estructural de proteínas eran incapaces de establecer estas relaciones estructurales.

Durante los últimos años se han desarrollado varias técnicas que permiten el análisis estructural comparativo de proteínas, obteniéndose aproximaciones interesantes. Sin embargo, estos análisis suelen ser muy costosos computacionalmente y suelen ofrecer unos resultados que enmascaran datos de interés biológico. La Figura IV.6 muestra gráficamente la superposición de tres moléculas con homología secuencial baja, como son el PCI (4cpaI), la toxina de escorpión (1agt) y la defensina HNP-3 (1dfnA). Aunque la superposición de los esqueletos polipeptídicos de estas proteínas dan valores relativamente altos, los carbonos α de las cisteínas se superponen con RMSD inferiores a 2 Å . Observando la Figura IV.6 podemos ver que se puede establecer fácilmente una relación estructural entre los esqueletos polipeptídicos de estas tres proteínas. Cuando se procede a realizar un análisis más profundo podemos ver como la región Q25-G35 del PCI, descrita como importante en la unión de esta molécula con el receptor del EGF (Blanco-Aparicio et al., 1998; Mas et al., 1998), tiene su equivalente en el lazo Y17-F29 de la defensina. Por otro lado el lazo D2-P8 de la defensina, descrito como importante para la dimerización de estas proteínas y consiguientemente relacionado con su función, tiene su equivalente con el lazo D20-C28 del veneno de escorpión (Hill et al., 1991; White et al., 1995). Las defensinas dimerizan formando unos canales que permeabilizan las membranas celulares produciendo así la muerte celular. Por su parte, el veneno de escorpión actúa uniéndose a los canales de K^+ (MacKinnon et al., 1998). Se ha sugerido que las toxinas de escorpión tienen un plegamiento CS$\alpha\beta$ basado en un cuerpo central $\alpha\beta$ y estabilizado por puentes de azufre, similar a las estructuras de ciertas proteínas de la familia de las *cystein-rich*, como son las defensinas de insectos. Estas analogías no han sido establecidas para defensinas de mamíferos como las analizadas aquí (defensin-HNP3). Sin embargo, todas estas proteínas

han podido ser analizadas conjuntamente gracias a la distribución topológica de sus puentes de azufre. Las superposiciones de estas moléculas desvelan otras similitudes estructurales como es la relación estructural de l*a región* N-terminal y el motivo CS$\alpha\beta$ del veneno de escorpión y la arquitectura del esqueleto del PCI, particularmente las regiones S7-I15 y P11-S19 de ellas. Relaciones estructurales más profundas de estas proteínas podrían permitir inferir teóricamente sobre la actividad funcional de las mismas.

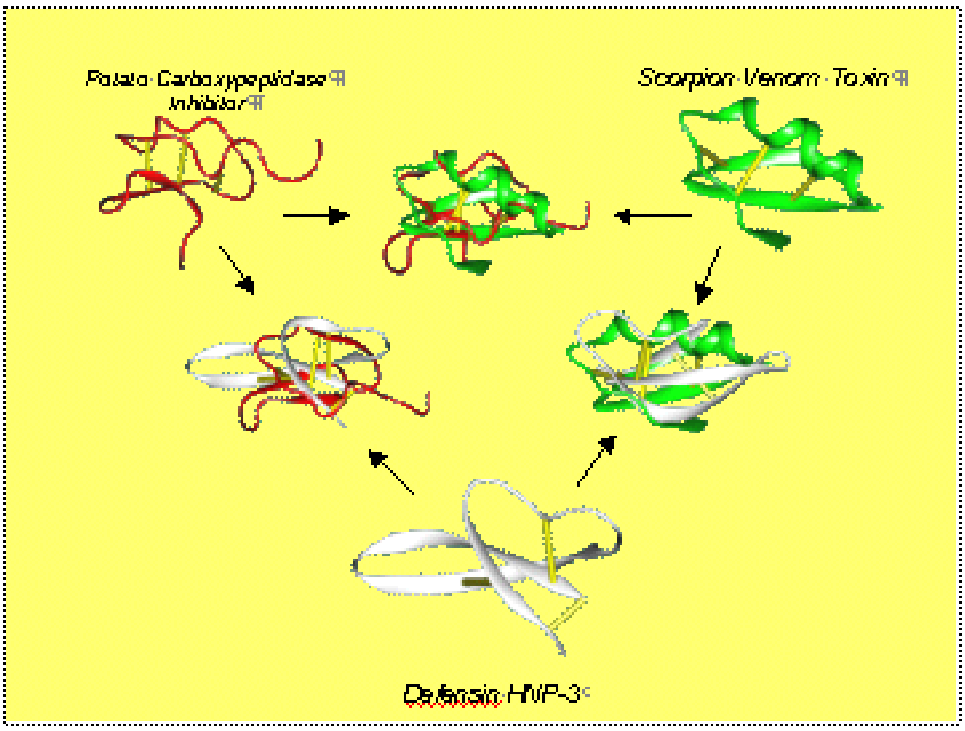


Figura IV.6 Esquema cíclico comparativo de la estructura de proteínas de distintas familias.

IV-B.5 Discusión y conclusiones

La comparación estructural de proteínas requiere la superposición de sus estructuras. Es casi infinito el número de formas de superponer la estructura de dos proteínas y el análisis que generaría sería extremadamente costoso para los ordenadores de hoy en día. En consecuencia, estas superposiciones se han realizado clásicamente a partir de alineamientos secuenciales o a partir de criterios extraídos desde las estructuras secundarias regulares. Debido a ello, la superposición de proteínas de baja homología secuencial y/o sin estructuras secundarias regulares comunes resulta muy complicado. En estos casos, la superposición depende del criterio subjetivo del investigador. Un importante grupo de proteínas ricas en puentes de azufre son de pequeño tamaño y con poca o ninguna estructura secundaria regular. Muchas de estas proteínas son de gran interés biotecnológico. Han sido descritas similitudes en el esqueleto peptídico y la topología de sus puentes de azufre entre algunas de ellas, como los venenos de escorpión, factores de crecimiento, inhibidores de proteasas o defensinas. Sin embargo, sus secuencias no presentan similitudes detectables, lo que las hace solamente comparables bajo puntos de vista estructurales (Rozwarski D.A., 1994). A su vez, muchas familias son descubiertas a partir de técnicas de comparación estructural (Holm & Sander, 1994; Murzin A.G., 1994; Orengo C.A. 1994). En este sentido, hemos observado cómo las superposiciones de puentes de azufre de proteínas con baja homología secuencial permite:

i) Obtener alineamientos del esqueleto peptídico. Estos alineamientos muestran, en algunos casos, interesantes propiedades como son:

i.a) El correcto alineamiento de estructuras secundarias regulares, observándose este efecto incluso entre proteínas de distintas famillias.

i.b) El correcto alineamiento de regiones características de esqueleto peptídico.

ii) Se observa que ciertas propiedades fisicoquímicas se conservan en el espacio. En algunos casos, donde la bibliografía es más amplia, vemos cómo estas posiciones corresponden a residuos que tienen un papel en el plegamiento, estabilidad o funcionalidad de las proteínas.

Podemos por tanto afirmar que la superposición de proteínas ricas en puentes de azufre ha sido útil en la caracterización de regiones y de residuos estructural y/o funcionalmente importantes. Pensamos, que para la comparación estructural de este tipo de proteínas, sobretodo cuando la homología secuencial es baja, es útil utilizar la topología de puentes de azufre.

Importancia biológica de este procedimiento

Como se ha podido ver en el análisis de la familia de los *EGF-like,* la superposición estructural que ofrece el KNOT-MATCH permite alinear en el espacio residuos estructural y funcionalmente importantes entre miembros de una familia. Algunas veces, estos alineamientos no coinciden con alineamientos secuenciales (más alejados del sentido funcional de las proteínas) y por tanto, quedan ocultos a los métodos clásicos. Un claro ejemplo del potencial que ofrece el alineamiento estructural frente a otro tipo de aproximación es la definición de las llamadas proteínas *T-knot* (Isaacs, 1995). Se han establecido dos posibles definiciones para un grupo de proteínas pequeñas, ricas en puentes de azufre. Por un lado aparece la definición que se basa en el orden secuencial de las uniones de las cisteínas para formar los puentes de azufre (Lin & Nussinov, 1995). Por otro lado, Harrison & Sternberg en 1996 definieron un nuevo patrón que se basaba en la distribución espacial de los puentes de azufre respecto de una hoja β (β-*cross*). El ya redundante ejemplo de la comparación entre el EGF y el PCI nos hace observar que ambos métodos son poco flexibles. Ambas proteínas, como se verá más ampliamente en capítulos posteriores, han sido relacionadas tanto estructural como funcionalmente. Por un lado, ya hemos comentado que el orden secuencial de las uniones de las cisteínas para formar los puentes de azufre es distinto entre estas proteínas, descartando el criterio

secuencial establecido. Por otro lado, la necesidad de la presencia de una hoja β es excesivamente estricto ya que, como ya ha sido comentado, el PCI tiene tan solo 5 residuos dispuestos en el espacio de manera que se pueda decir que están formando una hoja β. En los análisis realizados con el KNOT-MATCH hemos observado que esta distribución espacial de los puentes de azufre es común para proteínas como las feromonas. Éstas son proteínas pequeñas, de tamaño similar al PCI, pero que presentan una destacable hélice α y no una hoja β.

La técnica empleada es suficientemente flexible como para permitir establecer relaciones estructurales y funcionales entre este tipo de proteínas. La importancia estructural que tienen los puentes de azufre en estas proteínas, dado su pequeño tamaño, la carencia de estructuras secundarias regulares abundantes y el estrés ambiental al que se ven sometidas al realizar su función (generalmente se trata de proteínas con funciones extracelulares), hacen que los puentes de azufre cobren especial importancia. Un claro ejemplo de esto es la relación estructural establecida entre el EGF y el PCI confirmada experimentalmente (Blanco-Aparicio et al., 1998).

Es remarcable que muchas de las proteínas aquí tratadas: factores de crecimiento, defensinas, toxinas, feromonas, etc... son moléculas que suscitan un claro interés biotecnológico. El conocimiento más profundo de sus relaciones estructurales y funcionales, basado sobretodo en su análisis tridimensional, puede ser de gran ayuda para facilitarnos la comprensión de su mecanismo de funcionamiento, y en un proceso posterior, su rediseño o aplicación heteróloga. Una vez más cabe nombrar a la pareja EGF-PCI como proteínas que han sufrido, y siguen sufriendo, este proceso analítico.