



Secuencias proteicas cortas y sus estructuras secundarias: análisis informático e inferencia biológica

Diego Fernando Mejía Carmona

Universidad del Valle
Facultad de Salud
Escuela de Ciencias Básicas Médicas
Doctorado en Ciencias Biomédicas
Cali, Colombia, 2013

Secuencias proteicas cortas y sus estructuras secundarias: análisis informático e inferencia biológica

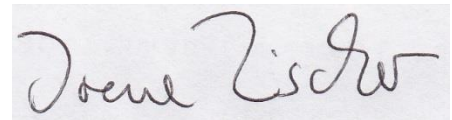
Diego Fernando Mejía Carmona

Tesis presentada como requisito parcial para optar al título de:
Doctor en Ciencias Biomédicas

Director: Irene Tischer, PhD.

Directora Grupo de investigación en
Bioinformática y Biocomputación
Universidad del Valle

Línea de Investigación:
Biología computacional de proteínas

A handwritten signature in dark ink, appearing to read 'Irene Tischer', on a light-colored rectangular background.

Firma del director

Irene Tischer, PhD

Grupo de Bioinformática y Biocomputación

Escuela de Ciencias de la Computación

Universidad del Valle

Santiago de Cali, Diciembre de 2013

Nota de aceptación

Firma del jurado

Firma del jurado

Firma del jurado

Santiago de Cali, Enero de 2013

DEDICATORIA

A mis padres,

Dolly Carmona y Alvaro Mejía

Todo un ejemplo de vida.

AGRADECIMIENTOS

A mi directora de tesis, Irene Tischer, líder del Grupo de Investigación en Bioinformática y Biocomputación de la Universidad del Valle, no sólo por su guía como directora de tesis, sino también por su paciencia, buen ánimo y todo el estímulo para seguir adelante con este trabajo.

A Cecilia Aguilar de Plata, profesora titular de la Facultad de Salud de la Universidad del Valle, por sus enseñanzas sobre proteínas, su amistad, su apoyo constante y su estímulo para terminar el trabajo, aún en momentos en que esto parecía imposible.

A Nilson Mossos, compañero doctorante del grupo de investigación en Bioinformática y Biología Computacional de Univalle, por producir la mayor parte del software para la generación de los primeros datos de secuencia y estructura, y su buena disposición para la ayuda en programación.

A Margot Edith Cuarán y Luis Garreta, compañeros doctorantes del grupo de investigación en Bioinformática y Biología Computacional de Univalle, por su solidaridad, y la producción de parte del software que permitió la obtención de los datos.

A Adalberto Sánchez, director del Postgrado en Ciencias Biomédicas, por su definitivo apoyo administrativo para que el presente trabajo se pudiera presentar en los plazos establecidos.

A Rachel Kramer – directora del PDB, por las respuestas a las consultas sobre ese banco de datos, por facilitar herramientas para la obtención de parte de los datos y por su interés en este trabajo.

A Paola Andrea Gómez Taborda, por su apoyo, su buen ánimo constante, su respaldo y su invaluable ayuda en la generación y edición de gráficos y tablas.

A Eric Gaucher, líder del Grupo de Biología Sintética de GeorgiaTech y a Ziming Zhao, hoy, postdoc en la Universidad de Yale, por enseñarme detalles y trucos de la reconstrucción ancestral.

A José Manuel Sánchez Ruiz, líder del grupo de investigación en Química Física e Ingeniería de Proteínas de la Universidad de Granada, por recibirme en su grupo y confiarme el rediseño de proteínas mediante reconstrucción ancestral.

A mi universidad, la universidad del Valle, por la formación académica que he recibido en ella; al Dr. Leonardo Fierro exdirector del postgrado en Ciencias Biomédicas, y a la Oficina de Relaciones Internacionales, en cabeza de Sandra Juliana Toro, por facilitar la pasantía de dos años para realizar parte de este trabajo en universidades de España y Estados Unidos.

A todas las personas que directa ó indirectamente participaron en este trabajo y en mi formación doctoral, y en especial a las personas del postgrado en Ciencias Biomédicas y el departamento de Ciencias Fisiológicas por hacer de estos años de doctorado, tan agradables y constructivos.

Presento mi agradecimiento póstumo al doctor Leonardo Lareo, investigador en proteínas de la Universidad Javeriana de Bogotá, por sus críticas constructivas, sugerencias y consejos sobre el manejo de algunos datos del presente trabajo.

TABLA DE CONTENIDO

INDICE DE FIGURAS	XI
INDICE DE TABLAS	XIV
PALABRAS CLAVE	XV
KEY WORDS	XVI
1.1 RESUMEN	XVIII
ABSTRACT	XIX
1.2 INTRODUCCIÓN	1
DESCRIPCIÓN DEL PROBLEMA Y ABORDAJE	1
<i>EL PROBLEMA</i>	1
Contexto	1
Problema Biológico y Computacional.....	4
1.3. JUSTIFICACIÓN.....	7
1.4. OBJETIVOS	9
1.4.1. OBJETIVOS GENERALES	9
1.4.2. OBJETIVOS ESPECÍFICOS.....	10
1.4.2.1. Seleccionar los conjuntos y subconjuntos de proteínas a ser estudiados, a partir del universo proteico representado en la totalidad de las proteínas del PDB, mediante la escogencia de una muestra representativa, con alta diversidad, baja redundancia y alta homogeneidad, en términos de la resolución de la estructura tridimensional.	10
1.4.2.2. Determinar la frecuencia de las secuencias cortas de residuos de aminoácidos en las proteínas, verificando si las secuencias cortas de aparición muy frecuente ó poco frecuente, presentan las mismas características cuando se analizan entre proteínas agrupadas según criterios como: su contenido de estructura secundaria, su localización celular, el organismo en el que se encontraron, el Phylum en que se encuentran, el tipo de Enzima según la Comisión de Enzimas, y familias de proteínas.	10
1.4.2.3. Analizar frecuencias de aminoácidos y tripletas en estructuras secundarias, Propensiones Chou y Fasman, Entropía conformacional e Indicador de Preferencia Estructural, buscando identificar propensiones características para proteínas agrupadas según los criterios antes mencionados.	10
1.4.2.4. Hacer uso de las frecuencias e IPE para Identificar y caracterizar zonas de una proteína que presentan mayor ó menor estabilidad para estructuras secundarias, y relacionar estas zonas con la funcionalidad del segmento donde ocurren.	10
1.4.2.5. Analizar las frecuencias de las las estructuras secundarias, en términos de las tripletas de residuos de aminoácidos, usando tanto los veinte aminoácidos, como un alfabeto reducido de propiedades fisicoquímicas de los aminoácidos.	10
1.4.2.6. Rediseñar proteínas mediante la reconstrucción de secuencias ancestrales, y evaluar experimentalmente las propiedades fisicoquímicas de las proteínas resucitadas.	10
1.5. MARCO TEÓRICO	11
<i>Cristalografía de rayos X como método para resolver la estructura 3D de una proteína.</i>	11
<i>Acerca del PDB</i>	11
La base de datos.....	11
El formato .pdb.....	12
<i>Sobre la asignación de estructura secundaria (DSS) a partir de las coordenadas 3D</i>	12
<i>Acerca de PFam</i>	14
<i>Sobre la estimación del orden en proteínas</i>	14
Entropía informacional usada como entropía estructural	14
<i>Sobre la preferencia estructural de secuencias cortas</i>	15
<i>Sobre Propensiones</i>	16
<i>Acerca de estructura secundaria de proteínas</i>	17
<i>Estructuras helicoidales</i>	17
<i>Hilos y hojas Beta</i>	17

<i>Giros Beta ó Giros-n</i>	18
<i>Dobléz</i>	18
<i>Estructuras al azar</i>	18
<i>Sobre estudios previos en secuencias cortas</i>	18
1.6. METODOLOGÍA	20
1.6.1. Selección de la muestra de proteínas para el análisis de la frecuencia y de la estructura secundaria a partir de los archivos .pdb	20
1.6.2. Asignación de estructura secundaria a partir de las coordenadas cartesianas de las posiciones de los átomos de cada proteína	21
1.6.3. Determinación de las frecuencias de aparición de aminoácidos individuales y tripletas, y relación entre estas frecuencias y la estructura secundaria	22
1.6.3. Análisis de la propensión de secuencias cortas de aminoácidos relacionadas con las secuencias estructurales.	22
1.6.4. Identificación y caracterización de zonas de mayor ó menor estabilidad respecto de la estructura secundaria	22
1.6.4. Sobre las tripletas, la generación de alfabetos reducidos, y las agrupaciones de las estructuras secundarias.	23
2.1. LA MUESTRA REPRESENTATIVA DEL PDB	31
2.2. LA MUESTRA DE PROTEÍNAS AGRUPADAS POR SU CONTENIDO DE ESTRUCTURA SECUNDARIA SEGÚN SCOP	32
2.3. LA MUESTRA DE PROTEÍNAS AGRUPADAS POR SU LOCALIZACIÓN CELULAR	34
2.4. LA MUESTRA DE PROTEÍNAS AGRUPADAS POR EL ORGANISMO DEL QUE PROVIENEN	34
2.5. LA MUESTRA DE PROTEÍNAS AGRUPADAS POR EL PHYLUM AL QUE PERTENECE EL ORGANISMO DEL QUE PROVIENEN	35
2.6. LA MUESTRA DE PROTEÍNAS AGRUPADAS POR EL TIPO DE ENZIMA SEGÚN LA NOMENCLATURA ENZIMÁTICA	36
2.7. LA MUESTRA DE PROTEÍNAS AGRUPADAS POR LA FAMILIA A LA QUE PERTENECEN, SEGÚN PFAM	36
2.8. VERIFICACIÓN DE LA REPRESENTATIVIDAD DE LA MUESTRA PDBGLOBAL	37
2.8.1. Comparación de las frecuencias de residuos entre PDBGlobal, PDBXRay, PDBUniverso y UniProt/SwissProt	37
2.8.2. Comparación entre PDBGlobal y PDBXRay en cuanto a las frecuencias de las estructuras secundarias	40
2.8.2.1. La participación de cada aminoácido en cada estructura secundaria	41
2.8.2.2. La distribución de las estructuras secundarias en cada uno de los aminoácidos	42
CONCLUSIONES CAPÍTULO 2 - SELECCIÓN DE LA MUESTRA	43
BIBLIOGRAFÍA CAPÍTULO 2 - SELECCIÓN DE LA MUESTRA	44
3.1. Frecuencias de residuos de aminoácidos en las proteínas agrupadas por los criterios seleccionados.	46
3.1.1.0. Frecuencias de aminoácidos en proteínas de la muestra PDBGlobal	47
3.1.1. Frecuencias de aminoácidos de las proteínas clasificadas por su composición de estructura secundaria, según SCOP	50
3.1.2. Frecuencias de aminoácidos de las proteínas clasificadas por su localización celular	56
3.1.3. Frecuencias de aminoácidos de las proteínas clasificadas por el organismo del que se aislaron	59
3.1.4. Frecuencias de aminoácidos de las proteínas clasificadas por Phylum	64
3.1.5. Frecuencias de aminoácidos de las proteínas clasificadas por tipo de enzima	66
3.1.6. Frecuencias de aminoácidos de las proteínas agrupadas por familia proteica, según PFam	69
3.1.7. Consolidado de Frecuencias de aminoácidos	70
3.2. Frecuencias de tripletas de residuos de aminoácidos en las proteínas agrupadas por los criterios seleccionados.	74
3.2.0. Frecuencias de tripletas en proteínas de la muestra PDBGlobal	74
3.2.1. Frecuencias de tripletas en proteínas clasificadas por su composición de estructura secundaria, según SCOP	85
3.2.2. Frecuencias de tripletas en proteínas clasificadas por su localización celular	87
3.2.3. Frecuencias de tripletas en proteínas clasificadas por Organismo	88
3.2.4. Frecuencias de tripletas en proteínas clasificadas por Phylum	89
3.2.5. Frecuencias de tripletas en proteínas clasificadas por Tipo de Enzima	90
3.2.6. Frecuencias de tripletas en proteínas clasificadas por familia proteica	91
CONCLUSIONES CAP. 3 - FRECUENCIAS DE AMINOÁCIDOS Y TRIPLETAS	93

BIBLIOGRAFÍA CAPÍTULO 3 – FRECUENCIAS DE AMINOÁCIDOS Y TRIPLETAS	96
CAPÍTULO 4. FRECUENCIAS DE AMINOÁCIDOS Y TRIPLETAS EN ESTRUCTURAS SECUNDARIAS	101
4.1. <i>Frecuencias de residuos de aminoácidos individuales en estructuras secundarias.</i>	101
4.1.1. Frecuencias de aminoácidos en las estructuras secundarias, en proteínas de PDBGlobal	102
4.1.2. Frecuencias de aminoácidos en las Estructuras secundarias, en los grupos de proteínas del presente estudio.	104
4.1.3. Frecuencias de aminoácidos en las Estructuras helicoidales, de las proteínas evaluadas en el presente estudio.	107
4.1.4. Frecuencias de aminoácidos en hilos y hojas Beta, de las proteínas evaluadas en el presente estudio	111
4.1.5. Frecuencias de aminoácidos en Giros Totales, de las proteínas evaluadas en el presente estudio.	114
4.1.6. Frecuencias de aminoácidos en Giros tipo Dobléz, de las proteínas evaluadas en el presente estudio.	115
4.1.7. Frecuencias de aminoácidos en Giros tipo Beta, de las proteínas evaluadas en el presente estudio	117
4.1.8. Frecuencias de aminoácidos en Random Coils “Giros al azar”, de las proteínas evaluadas en el presente estudio.	119
4.2. <i>Frecuencias de tripletas en las estructuras secundarias</i>	120
4.2.1. Frecuencias de tripletas en estructuras secundarias, en el PDBGlobal	120
CONCLUSIONES CAPÍTULO 4 - FRECUENCIAS DE LAS SECUENCIAS CORTAS EN LAS ESTRUCTURAS SECUNDARIAS	128
BIBLIOGRAFÍA CAPÍTULO 4 – FRECUENCIA DE SECUENCIAS CORTAS EN ESTRUCTURAS SECUNDARIAS.....	131
CAPÍTULO 5. TENDENCIA DE AMINOÁCIDOS A ADOPTAR ESTRUCTURAS SECUNDARIAS PARTICULARES ...	133
5.1. <i>Propensiones Chou y Fasman</i>	133
5.1.0. Propensiones de C&F en las proteínas de PDBGlobal.	133
5.1.0.4. Propensiones de Chou y Fasman, subclasificando los tipos de giros en dobléz, giro Beta, y giro al azar	138
5.1.1. Propensiones de C&F en proteínas agrupadas según SCOP.	139
5.1.2. Propensiones de C&F en las proteínas agrupadas por localización celular.....	141
5.1.3. Propensiones de C&F proteínas agrupadas por organismo	142
5.1.4. Propensiones de C&F en las proteínas agrupadas por Phylum.	144
5.1.5. Propensiones de C&F en las proteínas agrupadas por Tipo de enzima.....	146
5.1.6. Propensiones de C&F en las proteínas agrupadas por Familia proteica según PFam	147
5.2. <i>Comparaciones entre Propensiones de Chou y Fasman, Entropía Conformacional y el Indicador de Preferencia Estructural</i>	150
5.3. <i>Propensiones de Chou y Fasman, Entropías e IPE en tripletas de aminoácidos</i>	154
5.3.1. Las Propensiones en tripletas de aminoácidos.....	155
5.3.2. Las Entropías y el Indicador de Preferencia Estructural en tripletas de aminoácidos	160
CONCLUSIONES CAPÍTULO 5 - PROPENSIONES, ENTROPÍA E IPE PARA AMINOÁCIDOS Y TRIPLETAS	166
BIBLIOGRAFÍA CAPÍTULO 5 – PROPENSIONES C&F, ENTROPIA CONFORMACIONAL E INDICADOR DE PREFERENCIA ESTRUCTURAL	167
CAPÍTULO 6. REDISEÑO DE BMPs Y B-LACTAMASAS MEDIANTE RECONSTRUCCIÓN ANCESTRAL.....	171
TABLA DE CONTENIDO	172
6.1. INTRODUCCIÓN	173
6.1.1. <i>LA INGENIERÍA DE PROTEÍNAS</i>	173
6.3. OBJETIVOS.....	187
6.3.1. OBJETIVO GENERAL	187
6.3.2. OBJETIVOS ESPECÍFICOS.....	187
6.4. MATERIALES Y MÉTODOS.....	188
6.5. RESULTADOS Y DISCUSIÓN.....	190
6.6. CONCLUSIONES CAPÍTULO 6 - REDISEÑO DE BMPs Y B-LACTAMASAS MEDIANTE RECONSTRUCCIÓN ANCESTRAL.....	203
SIGUIENTES PASOS Y PERSPECTIVAS FUTURAS.....	204
CONCLUSIONES GENERALES	212
7.2 RECOMENDACIONES	215

7.3. PRODUCTOS Y PUBLICACIONES.....	216
7.3.1. <i>Del trabajo en proteínas del PDB</i>	216
7.3.2. <i>Del rediseño de proteínas por reconstrucción ancestral</i>	217

INDICE DE FIGURAS

Figura 2.8.1.1. Comparación entre las frecuencias de aparición de aminoácidos en proteínas de PDGlobal, PDBXRay, PDBUniversal y Uniprot/SwissProt.	39
Figura 2.8.1.2. Distribución de las frecuencias de aminoácidos en proteínas de PDGlobal.	39
Figura.1.8.2.1.1. Participación de cada uno de los 20 aminoácidos en las estructuras secundarias clásicas: hélices, hojas y giros, en las muestras PDGlobal, PDBXRay y PDBUniversal.	41
Como puede observarse en la figura 2.8.2.1.1., las frecuencias de aparición de los 20 residuos de aminoácido en las estructuras secundarias, son tan parecidas, que sólo se distinguen claramente, tres líneas, correspondientes a las estructuras helicoidales, hilos/hojas Beta y giros.	41
En la tabla 1.8.2.1.1., se muestran las frecuencias de aparición de los aminoácidos en las tres muestras, y debajo de todas, los valores de la prueba de t-student, por encima del umbral de 0.05, lo que confirma que las muestras provienen de la misma población y se pueden considerar estadísticamente iguales.	41
Figura 2.8.2.1.2. Participación de cada uno de los 20 aminoácidos en las estructuras secundarias DSSP-5, en la muestra PDGlobal.	42
Figura 2.8.2.2.1. La distribución de las estructuras secundarias en cada uno de los aminoácidos en la muestra PDGlobal.	43
Figura 3.1.0.1. Frecuencias de aminoácidos individuales en PDGlobal, organizadas en orden ascendente por frecuencia.	48
Figura 3.1.1.1. Magnitudes de las máximas diferencias entre las frecuencias del PDGlobal, y las frecuencias de las proteínas agrupadas según composición de estructura secundaria, según SCOP.	52
La línea negra en la diferencia cero, representa al PDGlobal.	52
Figura 3.1.1.2. Frecuencias de aparición de los veinte aminoácidos estándar en las proteínas agrupadas según su composición de estructura secundaria según la clasificación SCOP.	53
Figura 3.1.1.3. Frecuencias relativas de las proteínas agrupadas por composición de estructura secundaria según SCOP.	55
Figura 3.1.2.1. Frecuencias de aminoácidos individuales en las proteínas agrupadas por su localización celular, comparadas con las frecuencias de la muestra PDGlobal, y ordenadas ascendentemente.	57
Figura 3.1.3.1. Frecuencias de aminoácidos individuales en proteínas agrupadas por el organismo del que se aislaron.	60
Figura 3.1.3.2. Frecuencias relativas (normalizadas) de aminoácidos individuales en proteínas agrupadas por el organismo del que se aislaron.	60
Figura 3.1.4.1. Frecuencias de aminoácidos en proteínas agrupadas por Phylum, y comparadas con PDGlobal.	65
Figura 3.1.4.2. Frecuencias normalizadas de aminoácidos en proteínas agrupadas por Phylum, y comparadas con PDGlobal.	65
Figura 3.1.5.1. Frecuencias de aparición de los veinte aminoácidos en una muestra de proteínas enzimas, clasificadas según la Comisión de Enzimas.	67
Figura 3.1.5.2. Diferencias normalizadas como porcentaje de variación de las frecuencias de aminoácidos en las enzimas con respecto a las proteínas del PDGlobal.	67
Figura 3.1.6.1. Frecuencias de los aminoácidos individuales en las proteínas agrupadas por familia proteica según PFAM y comparadas con las frecuencias de la muestra PDGlobal.	69
Figura 3.1.6.2. Frecuencias de los aminoácidos individuales en las proteínas agrupadas por familia proteica según PFAM, normalizadas y comparadas con las frecuencias de la muestra PDGlobal.	69
Figura 3.1.7.2. Frecuencias de las propiedades fisicoquímicas seleccionadas en un alfabeto reducido de aminoácidos, en las proteínas en los 52 grupos de proteínas del presente trabajo.	71
Figura 3.1.7.3. Frecuencias normalizadas de los aminoácidos individuales en alfabeto reducido, en las proteínas en los 52 grupos de proteínas del presente trabajo.	72
Figura 3.2.0.1. Comparación entre frecuencias de tripletas observadas en el PDGlobal ordenadas de menor a mayor, y las frecuencias esperadas, para las mismas tripletas, calculadas con las frecuencias de los aminoácidos que las componen, obtenidas de la misma muestra de PDGlobal.	78

Figura 3.2.0.3. Comparación entre frecuencias observadas en tripletas de residuos de aminoácidos en el PDBGlobal ordenadas de menor a mayor, y las frecuencias observadas en la totalidad del PDB. En color azul, las frecuencias esperadas.....	81
Figura 3.2.0.4. Frecuencias de las 8000 tripletas, ordenada ascendentemente y de sus opuestas en secuencia.	81
En la figura 3.3.0.5. se puede notar que la hidropatía promedio ponderada es diferente para casi todas las tripletas, y que su valor puede ser importante en términos del aporte que hace a la hidropatía de la proteína. Esto muestra que se tiene una gama de ladrillos de construcción de proteínas, ó como el investigador Joji Otaki las llama, “tripletras constitutivas” (Otaki et al, 2005), con hidropatías variadas, y continuas normalizadas entre -1 y 1.....	82
Figura 3.2.0.5. Hidropatías (Kyte & Doolittle, 1982) de las tripletas. En rojo, el promedio ponderado (con respecto al aminoácido central (25%, 50%, 25%), y normalizado. En azul, el promedio simple y normalizado de las hidropatías de las tripletas. En negro, la frecuencia de aparición de cada una de las tripletas.	82
Figura 3.2.0.5. Frecuencia de tripletas por alfabeto reducido en el PDBGlobal. Arriba: tripletas completas. Abajo: tripletas reducidas.	84
Figura 3.2.1.1. Frecuencias de tripletas de residuos en proteínas agrupadas según SCOP.	85
Figura 3.2.2.1. Frecuencias de tripletas de residuos en proteínas agrupadas por su localización celular.....	87
Figura 3.2.3.1. Frecuencias de tripletas de residuos en proteínas agrupadas por Organismo.	89
Figura 3.2.4.1. Frecuencias de tripletas de residuos en proteínas agrupadas por Phylum.	90
Figura 3.2.5.1. Frecuencias de tripletas de residuos en proteínas agrupadas por tipo de Enzima.	91
Figura 3.2.6.1. Frecuencias de tripletas de residuos en proteínas agrupadas por familia proteica.	92
Figura 4.1.1.1. Frecuencia de aparición de aminoácidos en las estructuras secundarias del presente estudio en la muestra PDBGlobal.....	103
Figura 4.1.2.1. Frecuencia de aparición de las cinco estructuras secundarias en las proteínas del presente estudio (H, B, S, T, U). La suma de los tres tipos de giros, que llamamos “giros totales”, se encuentra en color gris.....	105
Figura 4.1.3.1. Frecuencia de aparición de aminoácidos en estructuras helicoidales, en las proteínas analizadas en el presente estudio.	108
Figura 4.1.4.1. Frecuencia de aparición de aminoácidos en hilos y hojas Beta, en las proteínas analizadas en el presente estudio.	112
Figura 4.1.5.1. Frecuencia de aparición de aminoácidos en Giros totales, en las proteínas analizadas en el presente estudio.	114
Figura 4.1.6.1. Frecuencia de aparición de aminoácidos en Giros tipo dobléz, en las proteínas analizadas en el presente estudio.	117
Figura 5.1.0.1. Comparación entre las frecuencias de los aminoácidos individuales entre la muestra PDBGlobal, y la muestra FrecLiterat, reportada por Costantini et al (2006).	134
Figura 5.1.0.2. Propensiones de C&F, hacia estructuras helicoidales, para aminoácidos individuales, en C&F original, la muestra PDBLit por Costantini et al, y la muestra PDBGlobal.	135
Figura 5.1.0.2. Propensiones de C&F, hacia hilos y hojas Beta, para aminoácidos individuales. Se presentan los valores originales de Chou y Fasman, la muestra PDBLit reportada por Costantini et al, y los de la muestra PDBGlobal.	136
Figura 5.1.0.4. Propensiones hacia la formación de giros (Pc), calculados originalmente por C&F, y las de las muestras PDBGlobal y PDBLit.....	137
Figura 5.1.0.5. Propensiones de Chou y Fasman calculadas para los tres tipos de giros, según DSSP.	138
Figura 5.1.1.1. Propensiones de C&F, en proteínas agrupadas según SCOP.	140
Figura 5.2.1. Chou & Fasman de aminoácidos en estructuras helicoidales, en proteínas agrupadas por su localización celular.	142
Figura 5.1.3.1. Propensiones Chou & Fasman de aminoácidos en proteínas agrupadas por Organismo.	143
Figura 5.1.4.1. Propensiones Chou & Fasman de aminoácidos en proteínas agrupadas por el Phylum al que pertenecen.	145
Figura 5.1.5.1. Propensiones Chou & Fasman de aminoácidos en proteínas agrupadas por el tipo de enzima, según la Comisión de Enzimas.....	146

Figura 5.1.6.1. Propensiones Chou & Fasman de aminoácidos en proteínas agrupadas por familia proteica según PFam.....	148
Figura 5.2.1. IPE, Entropías y Propensiones en los aminoácidos de PDBGlobal (para ajustar al rango entre cero y uno, se normalizaron las entropías y las propensiones).	151
Figura 5.2.2. IPE, Entropías y Propensiones en los aminoácidos en proteínas agrupadas por contenido de estructura secundaria según SCOP.	153
Figura 5.3.1. Propensiones hacia la formación de hélices (HHH), dobleces (SSS), y las transiciones de hélice a dobléz (HHS+HSS), y de dobléz a hélice (SSH+SHH).	156
Figura 5.3.2. Propensiones hacia la formación de hélices (HHH), giros Beta (TTT), y las transiciones de hélice a giro Beta (HHT+HTT), y de giro Beta a hélice (TTH+THH).	157
Figura 5.3.3. Propensiones hacia la formación de hélices (HHH), giros al azar (UUU), y las transiciones de hélice a dobléz (HHU+HUU), y de dobléz a hélice (UUH+UHH).	157
Figura 5.3.4. Propensiones hacia la formación de hélices (HHH), hilos y hojas Beta (BBB), y las transiciones de hélice a hilos (HHB+HBB), y de hilos a hélice (BBH+BHH).	158
Figura 5.3.5. Propensiones hacia la formación de hilos y hojas Beta (BBB), dobleces (SSS), y las transiciones de hilos/hojas a dobleces (BBS+BSS), y de dobleces a hilos (SSB+SBB).	158
Figura 5.3.6. Propensiones hacia la formación de hilos y hojas Beta (BBB), giros Beta (TTT), y las transiciones de hilos/hojas a dobleces (BBT+BTT), y de giros Beta a hilos (TTB+TBB).	159
Figura 5.3.7. Propensiones hacia la formación de hilos y hojas Beta (BBB), giros al azar (UUU), y las transiciones de hilos/hojas a giros aleatorios (BBU+BUU), y de giros Beta a hilos (UUB+UBB).	159
Figura 5.3.8. Propensiones hacia la formación de dobleces (SSS), giros Beta (TTT), y las transiciones de dobleces a giros Beta (SST+STT), y de giros Beta a dobléz (TTS+TSS).	160
Figura 5.3.8. Propensiones hacia la formación de giros Beta (TTT), giros al azar (UUU), y las transiciones de hilos/hojas a giros aleatorios (TTU+UUU), y de giros Beta a hilos (UUT+UTT).	160
Figura 5.3.2.1. Comparación entre el IPE calculado por tripletas de aminoácidos, y la Entropía Conformacional, también calculada por tripletas de aminoácidos.	161
Figura Comparación entre los valores de IPE de las tripletas que componen a diferentes grupos de proteínas (TGFP-BMP, Prión, PLA2, AFP, Archaea, GFP, Thermus, Thermotoga, Human, Small). En el eje X, están las tripletas que componen a cada una de las proteínas, y no son las mismas entre unas y otras proteínas. Se representan los valores de IPE de las tripletas que componen a varios tipos de proteínas.	165
Figura 5.1. Eventos celulares que ocurren después de la implantación de las BMP.	176
Figura 5.2. Secuencia de los 396 residuos de la preproteína de la BMP-2 humana. En color azul, la secuencia de los 114 residuos de la proteína madura activa.	180
Figura 5.3. Secuencia de los 114 residuos de la BMP-2 humana madura, con las Cisteínas en color, y los enlaces de disulfuro intra e intercatenarios.	181
Figura 5.4. Topología de la BMP-2 humana (generada en www.pdb.org).	181
Figura 6.5. Estructura tridimensional del dímero de BMP-2 humana (3BPM).	182
Figura 6.6. La estructura del complejo BMP-2–BRIAec. Vista lateral del lado próximo a la membrana.	183
Figura 6.7. Estructura del complejo BMP-2–BRIAec. Vista desde arriba, aproximadamente a lo largo de eje bilateral del complejo. En dorado y azul, monómeros de BMP-2. En verde, el receptor BRIAec. Se etiquetaron los elementos de estructura secundaria extremos de la cadena y los giros 1 y 3 del receptor (tomada de Kirsch, T. Et al. 2000).	183
Figura 5.8. Mecanismo de acción de las BMP-2.	185
Figura 6.9. Fragmento del alineamiento de las TGFB de los organismos seleccionados. Note la alta homogeneidad.	191
Figura 6.10. Árbol filogenético de la familia de proteínas parecidas a BMPs, de la superfamilia de las TGF-β.	192
Figura 6.11. Arbol filogenético mostrando algunos de los nodos ancestrales, de los cuales se reconstruyo el nodo 336 para resucitar en el laboratorio.	194
Figura 6.12. Detalle del árbol filogenético de las TGF-β, mostrando las ramas de las BMP-2 y las BMP-4. Se muestran los nodos ancestrales más cercanos a las proteínas BMP-2. Se reconstruyeron los nodos 340 y 341.	195
Figura 5.13. Mutaciones de los nodos ancestrales reconstruidos, con relación a la BMP-2 humana.	197

INDICE DE TABLAS

Tabla 2.1.1. Criterios tenidos en cuenta para la selección de las proteínas del presente estudio.	31
Tabla 2.2.1. Estadísticas de SCOP (Structural Classification of Proteins), versión 1.75.	33
Tabla 2.2.2. Criterios para la clasificación de proteínas por su contenido de estructura secundaria según SCOP, y número de proteínas seleccionadas.	33
Tabla 2.3.1. Proteínas seleccionadas para agruparlas por localización celular	34
Tabla 2.5.1. Proteínas seleccionadas y agrupadas por el Phylum al que pertenece el organismo del que se provienen	35
Tabla 2.6.1. Proteínas con actividad catalítica, clasificadas por tipo de enzima, de acuerdo con el “Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB)”	36
Tabla 2.7.1. Proteínas seleccionadas y agrupadas por familia proteica, según PFam	37
Tabla 2.8.1.1. Frecuencias de aminoácidos expresadas en porcentaje, en PDBGlobal, PDBXRay, PDBUniverso y Uniprot/SwissProt.	38
Tabla 2.8.2.1. Frecuencias de los aminoácidos en las estructuras secundarias DSSP-5 en PDBGlobal, PDBXRay y PDBUniverso, con los valores de la prueba t-student.....	41
Tabla 3.1.1. Alfabeto reducido para los veinte aminoácidos (aas), en un alfabeto de 9 letras, correspondiente a propiedades.	47
Tabla 3.1.0.1. Frecuencias de los veinte aminoácidos estándar en proteínas de la muestra PDBGlobal	48
Tabla 3.1.1.1. Frecuencias de aparición de los veinte aminoácidos estándar en las proteínas agrupadas según su composición de estructura secundaria según la clasificación SCOP.	50
Tabla 3.1.1.2. Diferencias máximas entre las frecuencias del PDBGlobal y las proteínas clasificadas por su contenido de estructura secundaria según SCOP.	52
Tabla 3.1.1.3. Diferencias normalizadas entre las frecuencias del grupo de referencia (PDBGlobal) y las proteínas clasificadas según SCOP.....	55
Tabla 3.1.2.1. Frecuencias de aminoácidos en proteínas agrupadas según su localización celular.	56
Tabla 3.1.3.1. Frecuencias de aparición de los veinte aminoácidos estándar en las proteínas agrupadas por el organismo del que se aislaron.....	59
Tabla 3.1.5.1. Frecuencias de aparición de los veinte aminoácidos en una muestra de proteínas del PDB clasificadas por tipo de enzima EC.....	67
Tabla 3.2.0.1. Frecuencias de aparición de las 20 tripletas menos frecuentes, en las proteínas de la muestra PDBGlobal.....	75
Tabla 3.2.0.2. Frecuencias de las tripletas menos frecuentes en PDBGlobal.	83
Tabla 3.2.0.2. Frecuencias de las tripletas menos frecuentes en PDBGlobal.	83
Tabla 4.1.1.1. Frecuencias de los veinte aminoácidos en las estructuras secundarias en el PDBGlobal.	102
Tabla 5.1.0.1. Propensiones de Chou y Fasman, clásicos, calculados con una muestra de 15 proteínas (Chou y Fasman, 1974a), los calculados con una muestra de más de 2000 proteínas (PDBSelect) (Costantini et al., 2006), y los calculados en nuestra muestra de más de 19000 proteínas (PDBGlobal). Pa se refiere a la propensión a formar hélices; Pb, a formar hilos/hojas Beta, y Pc, a formar giros.....	135
Tabla. 5.3.2.3. Umbrales de IPE y porcentaje de tripletas que constituyen diferentes tipos de proteínas....	164

PALABRAS CLAVE

Estructura secundaria, desorden en proteínas, comparación de secuencias, secuencias cortas, DSSP, propensión, entropía conformacional, estabilidad estructural, termoestabilidad, Indicador de preferencia estructural, estructuras al azar, proteínas intrínsecamente desestructuradas, secuencias camaleón.

Reconstrucción de secuencias ancestrales, Ingeniería de proteínas, resurrección de proteínas, promiscuidad enzimática.

KEY WORDS

Secondary structure, protein disorder, sequence comparison, short sequences, DSSP, propensity, conformational entropy, structural stability, thermostability, Indicator of structural preference, random coils, intrinsically unstructured proteins, chameleon sequences.

Ancestral sequence reconstruction, Protein engineering, protein resurrection, enzyme promiscuity.

CAPÍTULO 1. DESCRIPCIÓN DEL TRABAJO

1.1 RESUMEN

La vida como se conoce, no se podría concebir sin las funciones proteicas, que dependen de una estructura tridimensional particular para ser funcionales. La estructura a su vez, depende esencialmente de la secuencia de aminoácidos. Con esto, comprender la vida implica entonces comprender la forma como la secuencia proteica determina la estructura, y a su vez, la función.

El conocimiento sobre estructura y función de proteínas es fundamental en las ciencias de la vida, y da origen a campos nuevos como Ingeniería de Proteínas y Nanotecnología Biológica, industrialmente prometedores. La determinación experimental de la estructura es costosa en tiempo, dinero, infraestructura y recurso humano. Con los proyectos genoma y la gran cantidad de información que éstos producen, hay una creciente brecha entre la aparición de secuencias y el conocimiento de la estructura y función proteicas. Los esfuerzos se dirigen entonces a dilucidar computacionalmente la estructura y función proteicas, desde de la secuencia de sus aminoácidos.

En este trabajo analizamos datos experimentales de estructuras proteicas enfocándonos en aminoácidos individuales y tripletas de ellos, y su aparición en las estructuras secundarias de las proteínas almacenadas en el Banco de Datos de Estructuras de Proteínas (PDB).

A partir de las frecuencias de aminoácidos y tripletas, y de su aparición en estructuras secundarias, se dedujo información en términos de la composición de aminoácidos, estructura secundaria de diferentes grupos de proteínas, su estabilidad estructural y su tendencia a adquirir ó no, otra estructura secundaria, con lo que se identifican combinaciones de residuos que le confieren particularidades estructurales a una proteína, para ser usadas en el rediseño de proteínas.

Por otro lado, se reconstruyeron computacionalmente, Proteínas Morfogenéticas de Hueso (BMPs) y enzimas Beta-Lactamasas ancestrales, que luego se resucitaron en el laboratorio. Las Beta-Lactamasas ancestrales se caracterizaron en términos de su actividad enzimática y estabilidad térmica, encontrando que las enzimas reconstruidas presentan mayor promiscuidad de sustrato y mayor termoestabilidad, que sus equivalentes actuales. Las BMPs ancestrales reconstruidas muestran hasta el momento, mayor actividad en la regeneración ósea en cultivo y en huesos de animales pequeños, y se perfilan como candidatas para su uso en cirugías ortopédicas y maxilofaciales.

ABSTRACT

Life as we know it, couldn't be conceived without protein functions, these functions depend of a particular three dimensional structure. Structure, in turn, essentially depends of the aminoacids sequence. Understanding life therefore implies understanding of the way how a protein sequence determines structure, and the way how a structure conducts to the function of the molecule.

Knowledge of protein structure and function is fundamental in life sciences, and gives origin to new fields, industrially very promising, like Protein Engineering and Biological Nanotechnology. Experimental structure determination is costly in terms of time, money, infrastructure and human resources. With genome projects and the huge quantity of information they produce, there is a growing gap between solved genome sequences (and hence genes and protein sequences) and protein structure and function. Efforts have been addressed in elucidation of protein structure and function, through computational methods by using amino acid sequences.

In this work, we used protein bioinformatics tools to computationally analyze experimental data of protein structures, based on short amino acid sequences and their appearance in the secondary structures of the proteins stored in the Protein Data Bank (PDB).

From the frequencies of amino acids and residue triplets, and their appearance in secondary structures, we inferred information in terms of amino acid composition, secondary structure of different protein groups, their structural stability and their tendency to maintaining the same secondary structure or acquiring other, whereupon we inferred about residue combinations conferring structural particularities to a protein, to be used in protein engineering

On the other way, we reconstructed ancestral Bone Morphogenetic Proteins (BMPs), and β -Lactamase enzymes, to resurrect later in the lab. Ancestral β -Lactamases were characterized in terms of their enzymatic activity and thermal stability, finding that the reconstructed enzymes show enhanced substrate promiscuity and higher thermostability than their equivalents of present times. Reconstructed Ancestral BMPs showed until now, higher activity in bone regeneration in cell culture and in vivo, in animal bones, and are now candidates for using them in orthopaedic and maxillofacial surgery.

1.2 INTRODUCCIÓN

DESCRIPCIÓN DEL PROBLEMA Y ABORDAJE

EL PROBLEMA

Contexto

Considerando que las proteínas son las moléculas responsables de una enorme cantidad y variedad de funciones vitales para los seres vivos, que cada vez se usan más en el plano industrial y en aplicaciones de alta tecnología, y que en general, la estructura determina la función, es de interés hacer inferencia sobre las reglas que a partir de la secuencia y el ambiente, determinan la estructura. Por esta razón, en el presente trabajo, las proteínas se agrupan según diferentes criterios, en grupos claramente definidos, según los cuales, conjuntos de proteínas, presentan características similares. Es importante hacer estudios de las estructuras de las proteínas, para inferir sobre la forma como la secuencia y las características comunes a grupos de proteínas, interactúan para determinar la estructura.

De las investigaciones sobre las proteínas se puede derivar conocimiento útil no sólo en términos del desarrollo científico, sino también en términos de desarrollo aplicable en la medicina y en el campo tecnológico e industrial.

Las proteínas son cadenas lineales de aminoácidos unidos por enlaces covalentes conocidos como enlaces peptídicos. Estos enlaces confieren rigidez a la estructura, dado que se comportan como enlaces parcialmente dobles que pueden formar enlaces de hidrógeno, proveyendo estabilidad a la estructura, una vez plegada. Los aminoácidos encontrados en proteínas, son moléculas relativamente pequeñas, formadas por un grupo amino y un grupo carboxilo unidos a un carbono central conocido como Carbono Alfa.

Aunque hay más de 700 aminoácidos identificados, los que forman las proteínas y que están codificados a nivel de los genes, son sólo unos cuantos (22 identificados hasta ahora). Estos aminoácidos se distinguen unos de otros por la cadena lateral unida al Carbono Alfa. Algunas cadenas son hidrofílicas, otras hidrofóbicas; algunas son de tamaño muy pequeño, otras más grandes; unas son muy reactivas, otras no. Por esto, la cadena lateral de los aminoácidos les da características particulares a las secuencias proteicas, determinando la forma como la cadena de aminoácidos interactúa con su microambiente y se pliega dando como resultado su estructura y finalmente su función. Dicho de otra manera, la composición de aminoácidos y la forma como están combinados en la secuencia proteica, determina el plegamiento de la proteína. Pero aún

más importante, es que el plegamiento de una proteína conduce a la estructura tridimensional, que determina su función.

El plegamiento depende entonces de diferentes factores: la concentración de sales y cargas en el medio que rodea a la proteína, de la existencia de agua en el entorno, y de la interacción de las cadenas laterales de los aminoácidos con el medio. Sin embargo, en general se ha considerado cada aminoácido y su cadena lateral, individualmente en cuanto a su aporte a la estructura. En el presente trabajo, se considera a cada aminoácido de manera individual, y junto con sus vecinos en la cadena, de manera que se pueden tener en cuenta las interacciones entre ellos. Mediante el estudio del efecto que ejercen las diferentes combinaciones de aminoácidos sobre la estructura, se pueden deducir reglas que contribuyan a explicar el plegamiento y el comportamiento funcional de una proteína.

Se han desarrollado muchos métodos experimentales para el estudio de las proteínas y de los genes que las determinan, pero es claro que estudiar todas las proteínas de un organismo mediante este enfoque sería una labor extremadamente consumidora de tiempo y recursos, sobre todo considerando que por ejemplo, el ser humano tiene alrededor de 27.000 genes, que pueden producir más de un millón de proteínas diferentes (Chuan, 2012).

El estudio de los genes es mucho más económico que el de las proteínas. Esto es evidente en el hecho de que actualmente, la mayor base de datos de secuencias de proteínas, SwissProt-TrEMBL (<http://expasy.org/sprot/> Apweiler, et al, 2004) en su versión 2013.10 del 16 de Octubre de 2013, cuenta con 541.561 secuencias proteicas, mientras que en banco de datos de estructuras de proteínas, PDB (<http://www.pdb.org/pdb/statistics/holdings.do>), en su versión de Noviembre 5 de 2013 cuenta con 88.203 proteínas con estructura tridimensional resuelta. A la fecha, se han resuelto experimentalmente las secuencias de unos 168'335.396 genes (NCBI-GenBank release 198.0, Oct 15, 2013, <ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt>), mientras que sólo se han resuelto las estructuras de unas decenas de miles de proteínas; la brecha entre las proteínas cuya secuencia se resuelve y las proteínas cuya estructura se resuelve, es cada vez mayor.

Sin embargo, esta brecha puede disminuirse significativamente mediante el uso de herramientas computacionales que permitan identificar la estructura y la función de las proteínas determinadas por los genes secuenciados por los proyectos genoma y los diferentes grupos de investigación en genética, ahorrando así el costo de la determinación experimental de la estructura y la función de cada proteína.

Por esto, el enfoque computacional se considera tan prometedor, y está además al alcance de los países del tercer mundo, dadas las diferencias económicas que existen entre el costo de la infraestructura para el estudio experimental, y el costo de la infraestructura para los estudios computacionales (ICGEB, 2001).

Ha habido muchos esfuerzos por deducir las reglas que determinan el plegamiento de las proteínas y su subsecuente influencia en la función, pero aún no se ha logrado predecir, con satisfactoria precisión, la estructura y la función de una proteína en particular, y mucho menos deducir reglas universales aplicables al estudio de la generalidad de las proteínas.

Hay varios niveles estructurales en proteínas, desde primaria (la secuencia de aminoácidos), hasta cuaternaria (la organización de múltiples proteínas en proteínas multiméricas); incluso, se considera un quinto nivel, la agregación de múltiples complejos proteicos para cumplir una función específica en conjunto, como en el transcriptoma. Entre estos niveles estructurales, se encuentra la estructura secundaria (plegamientos locales dados por residuos de aminoácidos contiguos en la secuencia) y terciaria (plegamientos globales entre aminoácidos lejanos). La estructura secundaria, parece ser un intermediario importante en el plegamiento de la estructura proteica, al punto que se considera como un precursor de la estructura terciaria de la proteína (Gong y Rose, 2005), sobre todo en los modelos de nucleación en el plegamiento proteico (Wetlaufer, 1973; Abkevich et al., 1994; Dill et al., 1995; Djikaev y Ruckenstein, 2007).

Estudiar la estructura secundaria es un punto crítico para el avance en la comprensión y predicción de la estructura tridimensional de estas moléculas. Hay evidencia de que este nivel estructural es determinante para el plegamiento final de la proteína, y es responsable en buena parte, de los movimientos locales en la proteína, que participan en el desempeño de la función proteica (Myers y Oas, 2001).

Las proteínas realizan muchas funciones vitales y esto puede usarse en la medicina para corregir, desde el nivel molecular, problemas bioquímicos que causan enfermedades (en campos como Medicina molecular, Terapia Génica, Diseño de Medicamentos, Biofarmacología y Terapia Enzimática); también puede aprovecharse de la capacidad catalítica de muchas proteínas usándolas en la industria química y biotecnológica (Enzimología e Ingeniería de Enzimas), algunas proteínas pueden participar en la producción de nanomáquinas (Nanotecnología proteica), en el diseño de eficientes memorias de computador (Majima, 2008), y además se pueden diseñar proteínas con características deseables como mayor termoestabilidad ó promiscuidad ampliada (Risso, et al., 2012) nuevas funciones proteicas, deseables para otros procesos industriales (Ingeniería de proteínas). El estudio y conocimiento de las proteínas ofrece una oportunidad no sólo para el avance científico, sino también el desarrollo de industria de base tecnológica.

El estudio de las proteínas es de interés global y sus aplicaciones pueden ser de vital importancia en el desarrollo biotecnológico del país. Hay estudios sobre estructuras proteicas desde hace alrededor de treinta años en el mundo, aunque el enfoque computacional se usa masivamente sólo en años más recientes; en Colombia se trabaja sólo desde hace unos pocos años desde la perspectiva de los enfoques computacionales. Entre los grupos colombianos de estudios bioinformáticos en proteínas se encuentran: el grupo de Bioquímica Computacional de la Universidad Javeriana de Bogotá, que combina enfoques experimentales con los computacionales, el grupo de Biotecnología y Bioinformática de la Universidad Nacional de Colombia y el Grupo de

Bioinformática y BioComputación de la Universidad del Valle, que se enfoca en modelos computacionales para el estudio de proteínas y de genomas. (Al parecer, la mayoría de grupos colombianos dedicados a bioinformática o biología computacional, han puesto su esfuerzo principalmente en el estudio de genes y genomas).

Este proyecto se desarrolló como parte de las actividades de investigación del grupo de Bioinformática y BioComputación de la Universidad del Valle (en Ciencias de la Computación), en la modalidad de tesis doctoral dentro del Postgrado en Ciencias Biomédicas de la misma universidad, y hasta el momento el enfoque y sus resultados han producido dos capítulos de libro (Mejía y Tischer, 2007; y Mossos, Mejía-Carmona, Tischer, 2014), y se ha presentado en eventos como La semana de la Ingeniería, Universidad del Valle 2009; Segundo seminario internacional sobre Genómica, Proteómica, Bioinformática y Biología de Sistemas, Universidad del Cauca, Popayán, 2007; y Segundo congreso colombiano de Bioinformática y Biología Computacional, BIOS, Centro de Bioinformática y Biología Computacional, 2013).

Problema Biológico y Computacional

Los resultados de los trabajos experimentales realizados en la solución de la estructura tridimensional de las proteínas se almacenan generalmente en un banco público de datos conocido como PDB (Protein Data Bank). Se almacenan las estructuras tridimensionales en forma de coordenadas cartesianas para cada uno de los átomos diferentes de los hidrógenos. Además se encuentra la información básica de la proteína como nombre de la proteína, especie biológica donde se localiza, localización celular, función, nombre de la proteína, referencia bibliográfica, etc. Esta información se almacena en forma de archivos de texto plano (con extensión .pdb), lo que permite extraer su información fácilmente.

El PDB sin embargo, es un reservorio de datos que requiere de un cuidadoso proceso de selección. Desde la versión 3.1 - 2006 del PDB, se ha hecho un exhaustivo trabajo de remediación de los datos, corrigiendo inconsistencias, completando anotaciones, y depurando la información para dar un PDB mucho más confiable y seguro que antes (Henrick *et al.*, 2007).

Los archivos .pdb, que son archivos de texto plano que contienen la información para cada proteína, no dicen cuál es la estructura secundaria de la misma. Esto debe deducirse a partir de las coordenadas cartesianas de los átomos, en lo que se conoce como “asignación de estructura secundaria”. Aunque hay varios algoritmos y enfoques, el algoritmo DSSP (Dictionary of Secondary Structure of Proteins) desarrollado por investigadores del Instituto Max Planck (Kabsch y Sander 1983), es el comúnmente usado, y el “estándar de oro” para este aspecto (Martin *et al*, 2005). Este algoritmo basa la asignación de estructura secundaria en los patrones de formación de enlaces de hidrógeno en la molécula. El DSSP (Kabsch y Sander 1983), asigna ocho tipos de estructuras secundarias, así:

1. **Giros Beta y giros-n:** un patrón básico de estructura secundaria. Un residuo de aminoácido se asigna como perteneciente a un giro si hay un enlace de hidrógeno entre el CO(i) y el NH(i+n), con n=3,4, ó 5). Hay cuatro tipos de giro :
 - a. **Giro tipo I:** los ángulos diédricos son (-60,-30) y (-90, 0) para los residuos i+1 e i+2, respectivamente.
 - b. **Giro tipo II:** los ángulos diédricos son (-60, 120) y (80, 0) para los residuos i+1 e i+2 respectivamente.
 - c. **Giro tipo III:** Los ángulos son (-60, -30) y (-60, -30) para los residuos i+1 e i+2 respectivamente. Este giro es la unidad más simple de una hélice 3_{10} .
 - d. **Giro Gamma:** Es un tipo especial de giro y tiene un enlace de hidrógeno entre el CO del residuo i y el NH del residuo i+2. Los ángulos diédricos para el residuo i+1 son (70, -60) y (-70, 60) si es un giro Gamma inverso.
2. **Hélice Alfa:** definida como al menos dos giros consecutivos, y se asigna al residuo i, cuando se identifica un enlace de H entre el aminoácido i y el i+3.
3. **Hélice 3_{10} :** se asigna al residuo i, cuando hay un enlace de H entre i e i+2.
4. **Hélice Pi:** con enlace de H entre i e i+4.
5. **Puente Beta:** Al menos dos conjuntos contiguos de tres residuos cada uno, (i-1, i, i+1 y j-1, j, j+1). Se identifican dos tipos de puentes:
 - a. **Puente paralelo (i,j):** con enlaces de H: (i-1, j) y (j, i+1), ó (j-1, i) y (i, j+1).
 - b. **Puente antiparalelo (i,j):** con enlaces de H: (i, j) y (j, i) ó (i-1, j+1) y (j-1, i+1)
6. **Escalera Beta y Hoja Beta:**
 - a. **Escalera Beta:** Se define como un conjunto de puentes consecutivos de idéntico tipo (paralelo ó antiparalelo)
 - b. **Hoja Beta:** Es un conjunto de dos o más escaleras conectadas por residuos compartidos.
7. **Dobléz:** Es una región con alta curvatura. Para un residuo i, la curvatura se cuantifica como el ángulo del esqueleto de (i-2, i-1, i) y de (i, i+1, i+2). Para un dobléz en I, se requiere una curvatura de al menos 70 grados:

$$\text{Dobléz (i)} = \text{ángulo } \{C\text{-}\alpha(i)\sim C\text{-}\alpha(i-2), C\text{-}\alpha(i+2)\sim C\text{-}\alpha(i)\} > 70$$
8. **Estructura al azar:** Es una región donde el residuo no puede asignarse a ninguna de las anteriores estructuras secundarias.

Muchos de los enfoques para el estudio computacional de proteínas se basan en la premisa de que la secuencia de aminoácidos determina la estructura de la proteína, como propuso Christian Anfinsen con su trabajo sobre la enzima ribonucleasa de bovinos en 1954 (Anfinsen et al., 1954; Epstein et al., 1971). En años más recientes, algunos de los esfuerzos se han enfocado en la idea de encontrar una potencial tendencia intrínseca de cada residuo de aminoácidos a formar una estructura secundaria particular (conocida como propensión) (Chou y Fasman, 1978). Este fue posiblemente el inicio de los intentos de predicción de estructura de proteínas. Posteriormente, Garnier et al., (1978) propusieron otro algoritmo de predicción de estructura secundaria. Sin embargo estos enfoques alcanzaron bajos índices de predictibilidad (calculad como el número de aminoácidos correctamente predichos con respecto al número total de aminoácidos), según cita el

mismo Gerald Fasman en su libro “Prediction of Protein Structure and the Principles of Protein Conformation” (1989).

Las propensiones de Chou y Fasman han sido reevaluadas en varias ocasiones (Kynğäs y Valjakka, 1997; Lareo y Acevedo, 2005); en cuanto a esto, al inicio de este trabajo, publicamos la propuesta de un indicador de la preferencia estructural, que cuantifica la tendencia de una tripleta de aminoácidos, a formar estructuras secundarias estables (Mejía y Tischer, 2007). Sin embargo, es posible que la baja predictibilidad de los enfoques estadísticos que usan propensiones se deba a que consideran a cada aminoácido individualmente y no tienen en cuenta el contexto (los aminoácidos vecinos en la secuencia). En el trabajo de Lareo y Acevedo (2005), y el de Otaki et al. (2005), consideran secuencias cortas de aminoácidos. Por ello consideramos conveniente hacer un estudio de cada aminoácido junto con sus vecinos contiguos en una secuencia proteica. Es importante aclarar que a lo largo de este documento, se usan indistintamente los términos residuo de aminoácido, residuo y aminoácido, para referirse a los aminoácidos que forman parte de una cadena peptídica en una proteína.

El presente enfoque ofrece una alternativa coherente para la predicción local de la estructura de las proteínas, la identificación de proteínas y/o regiones de proteínas estructuralmente estables, la comparación de proteínas clasificadas según sus características, y la identificación de proteínas potencialmente relacionadas con enfermedad debido a su potencial tendencia a plegarse incorrectamente.

1.3. JUSTIFICACIÓN

Los sistemas biológicos son complejos en términos de su función y sus interacciones. Presentan múltiples aspectos que todavía no están comprendidos. Uno de los cuestionamientos centrales de la Biología se refiere a la forma como los genes determinan una función celular; esto implica conocer cómo la secuencia del gen (y por tanto, la secuencia de aminoácidos en la proteína), determina la estructura tridimensional de la misma, y cómo ésta estructura determina la función. Así que resolver cómo la secuencia determina la estructura, permitirá avances en la comprensión de enfermedades, envejecimiento, y el desarrollo celular, y a través de ella, la modificación de estos procesos con fines terapéuticos, el desarrollo de medicamentos, la producción de nuevas proteínas con funciones adecuadas para procesos de interés humano, y el desarrollo de máquinas moleculares con los mismos fines.

De muchas proteínas se conoce la estructura y la función, pero también hay una enorme y creciente cantidad de proteínas con secuencia conocida, pero estructura y función desconocidas. Por eso surge la necesidad de desarrollar métodos de análisis experimental para resolver la estructura proteica y métodos computacionales para predecir la estructura y función de proteínas no resueltas experimentalmente.

La predicción de la estructura de las proteínas en el momento actual alcanza una precisión poco mayor del 70%. Este valor es posible gracias a que la secuencia de aminoácidos tiene inherentemente la información estructural (propensión) (Anfinsen, 1971; Chou y Fasman, 1978). En este proyecto se extiende la definición de propensión de un aminoácido individual, a secuencias cortas de aminoácidos, y además, se extiende a diferentes tipos de giros en proteínas, discriminando Giros tipo Beta, con enlaces de H estabilizantes; Doble, sin enlace de H; y Giros al azar, conocidos también como “random coils”, cuyo término usaremos indistintamente a lo largo de este documento.

La dificultad de mejorar la predicción se debe en parte a la existencia de zonas desordenadas en las proteínas. Sin embargo, estas zonas son en muchos casos responsables de la función. El presente proyecto busca identificar aquellas zonas de mayor orden y desorden en secuencias proteicas.

Aunque hay iniciativas y esfuerzos internacionales para comprender las proteínas (como “The Protein Structure Initiative” en www.structuralgenomics.org), aún hay un vasto campo para explorar. Se desconoce parcialmente cómo una proteína puede alcanzar su estructura tridimensional nativa, y cómo puede cambiar esta estructura para adquirir una nueva función, ó perder la función original, resultando a veces en un evento causante de una enfermedad. En las estructuras proteicas se ha encontrado un alto grado de orden; sin embargo, también se han encontrado zonas desordenadas; actualmente se discute el valor del desorden desde el punto de vista funcional, pues algunos autores han identificado una relación entre zonas desordenadas y

zonas funcionales en las proteínas (Lobley et al., 2007). En el presente trabajo, se propondrá una metodología para la identificación de zonas con alto grado de estabilidad estructural, posiblemente responsables del andamiaje de la proteína, y zonas con bajo grado de estabilidad, y a partir de allí, de zonas desordenadas que pueden presentar mayor tendencia a cambiar su estructura, facilitando la ocurrencia de un plegamiento diferente en la proteína, que podría potencialmente modificar la función. Esto permitiría identificar en un genoma, las proteínas candidatas a producir enfermedades por errores ó modificaciones en el plegamiento.

Este proyecto contribuyó al estudio computacional de proteínas del grupo de investigación en Bioinformática y Biología Computacional de la Universidad del Valle, aporta conocimiento al análisis de información masiva del Banco de Datos de Estructuras Proteicas, PDB, y propone herramientas para el rediseño de proteínas.

1.4. OBJETIVOS

1.4.1. OBJETIVOS GENERALES

1.4.1.1. Analizar las frecuencias de aminoácidos individuales y combinaciones de tres residuos (tripletas), junto con su estructura secundaria, en proteínas con estructura tridimensional resuelta, e inferir acerca de su potencial predictivo para la estructura y características estructurales de las proteínas.

1.4.1.2. Rediseñar enzimas β -Lactamasas y Proteínas Morfogenéticas de hueso mediante la reconstrucción de secuencias ancestrales y evaluar las propiedades fisicoquímicas de las proteínas rediseñadas.

1.4.2. OBJETIVOS ESPECÍFICOS

1.4.2.1. Seleccionar los conjuntos y subconjuntos de proteínas a ser estudiados, a partir del universo proteico representado en la totalidad de las proteínas del PDB, mediante la escogencia de una muestra representativa, con alta diversidad, baja redundancia y alta homogeneidad, en términos de la resolución de la estructura tridimensional.

1.4.2.2. Determinar la frecuencia de las secuencias cortas de residuos de aminoácidos en las proteínas, verificando si las secuencias cortas de aparición muy frecuente ó poco frecuente, presentan las mismas características cuando se analizan entre proteínas agrupadas según criterios como: su contenido de estructura secundaria, su localización celular, el organismo en el que se encontraron, el Phylum en que se encuentran, el tipo de Enzima según la Comisión de Enzimas, y familias de proteínas.

1.4.2.3. Analizar frecuencias de aminoácidos y tripletas en estructuras secundarias, Propensiones Chou y Fasman, Entropía conformacional e Indicador de Preferencia Estructural, buscando identificar propensiones características para proteínas agrupadas según los criterios antes mencionados.

1.4.2.4. Hacer uso de las frecuencias e IPE para Identificar y caracterizar zonas de una proteína que presentan mayor ó menor estabilidad para estructuras secundarias, y relacionar estas zonas con la funcionalidad del segmento donde ocurren.

1.4.2.5. Análizar las frecuencias de las las estructuras secundarias, en términos de las tripletas de residuos de aminoácidos, usando tanto los veinte aminoácidos, como un alfabeto reducido de propiedades fisicoquímicas de los aminoácidos.

1.4.2.6. Rediseñar proteínas mediante la reconstrucción de secuencias ancestrales, y evaluar experimentalmente las propiedades fisicoquímicas de las proteínas resucitadas.

1.5. MARCO TEÓRICO

Cristalografía de rayos X como método para resolver la estructura 3D de una proteína.

En 1958, los investigadores John Kendrew y Max Perutz en forma independiente, resolvieron la estructura tridimensional de dos proteínas: la mioglobina y la hemoglobina. Para ello usaron una técnica experimental relativamente novedosa: la cristalografía de rayos X (Kendrew et al., 1958; Muirhead y Perutz, 1958). Por estos trabajos, compartieron el premio Nóbel de Química en 1962. Con esto, la metodología de Cristalografía de rayos X demostró que era una técnica poderosa para dilucidar estructuras tridimensionales de grandes moléculas.

La cristalografía de rayos X es hasta el momento, la metodología de primera opción para la dilucidación de la estructura de moléculas, tanto biológicas como no biológicas; al punto que de las 88203 proteínas resueltas hoy (Noviembre 8 de 2013), la mayoría (78612) se resolvieron por esta técnica (PDB Statistics: <http://www.rcsb.org/pdb/statistics/holdings.do>).

En términos generales, la técnica consiste en purificar en un alto grado la molécula cuya estructura se quiere resolver, para luego cristalizarla, y bombardear el cristal con Rayos X, con el fin de que estos, luego de desviarse por los átomos de la proteína en el cristal, impacten posteriormente una placa fotográfica. En esta placa, se obtienen puntos debidos al impacto de los rayos X, y la distribución espacial de los puntos en la placa que refleja el grado de desviación de los rayos X al atravesar las moléculas atrapadas en el cristal con la molécula a resolver. Esta metodología ha recibido numerosas mejoras en los últimos años, en especial, la cristalización de las proteínas, que usualmente es un cuello de botella (Giacovazzo, et al, 2002).

Acerca del PDB

La base de datos

El Banco de Datos de Estructuras de Proteínas (PDB por Protein Data Bank) es una base de datos mundial con la información de la estructura tridimensional de macromoléculas biológicas, en especial, las proteínas. Se estableció en los Laboratorios Nacionales Brookhaven (BNL por Brookhaven National Laboratories) en 1971 como un archivo para las estructuras obtenidas por cristalografía de rayos X de macromoléculas biológicas (Meyer, 1997, PDB, 1971). Empezó con siete estructuras y aumentó cada año, hasta que en los años 80 creció dramáticamente debido a mejoras en las técnicas cristalográficas, en resolución de estructuras por resonancia magnética nuclear y cambios en la concepción de la comunidad científica acerca del compartir los datos (Berman et al., 2000). Hoy en día, el PDB está bajo la responsabilidad de la Colaboración Investigativa para la Bioinformática Estructural (RCSB por Research Collaboratory for Structural Bioinformatics), que es un consorcio sin ánimo de lucro dedicado a mejorar nuestro

entendimiento de la función de sistemas biológicos mediante el estudio de la estructura tridimensional de macromoléculas biológicas (Meyer, 1997). Al PDB puede accederse públicamente vía internet en <http://home.rcsb.org>.

Actualmente, el PDB se ha convertido en el principal repositorio de macromoléculas biológicas con estructuras tridimensionales resueltas. Con más de 45000 estructuras, el PDB está disponible en línea como una importante fuente de información para la química, la biología estructural, la biología computacional, farmacología y la educación. Durante los últimos años, los miembros de la organización internacional que anota y libera los datos al PDB (la wwPDB : worldwide Protein Data Bank), han hecho un enorme esfuerzo por remediar y homogenizar la información almacenada en el PDB (Henrick et al., 2007). wwPDB es un esfuerzo colaborativo entre el Instituto Europeo de Bioinformática (MSD-EBI por European Bioinformatics Institute), El Banco de Proteínas del Japón (PDBj por Protein Data Bank – Japan), y el Banco de Datos de Resonancia Magnética Biológica (BRMB por Biological Magnetic Resonance Data Bank – USA). Se puede encontrar más información acerca del PDB en www.pdb.org/pdb/static.do?p=general_information/about_pdb/index.html.

Virtualmente, cada estructura almacenada en el PDB ha sido ampliada y actualizada para mejorar la información. A mediados del 2007 se actualizó el PDB para incorporar esta información revisada y mejorada (Henrick et al., 2007, PDB Annual Report, 2007). Ahora, el PDB se ha hecho mucho más uniforme en cuanto a la información que contiene (PDB Annual Report 2007). Hoy en día, el PDB provee un sitio que evoluciona permanentemente y que ofrece una plataforma para el estudio de macromoléculas biológicas y sus relaciones con la secuencia, la función y la enfermedad.

El formato .pdb

Cada proteína con estructura 3D resuelta se almacena como un archivo de texto plano con extensión .pdb. Este formato fue propuesto y establecido inicialmente como parte de la iniciativa de poner los datos en una base de público acceso en la red mundial de internet (Berman *et al.*, 2003).

Los datos contenidos en un archivo .pdb incluyen información general sobre la proteína, la secuencia de residuos de aminoácidos, comentarios de los autores sobre la localización celular y función de la proteína, coordenadas atómicas de la estructura tridimensional de la proteína, citas bibliográficas, estructura primaria de la proteína, e información sobre los datos experimentales de la elucidación de la estructura tridimensional de la proteína (Atomic coordinate entry format description, 2007).

Sobre la asignación de estructura secundaria (DSS) a partir de las coordenadas 3D

Es importante categorizar los elementos de estructura secundaria, pues este es un paso importante para comprender la estructura de las proteínas, su función y las interacciones necesarias para cumplir la función. Hay varios algoritmos que hacen la asignación de estructura

secundaria de proteínas, a partir de las coordenadas 3D de los átomos consignados en los archivos pdb.

El primer algoritmo fue propuesto por los investigadores Levitt y Greer (1977), y se basaba principalmente en los ángulos de torsión entre los carbonos alfa. Pocos años después, investigadores del instituto Max Planck, Kabsch y Sander (1983), desarrollaron un algoritmo conocido como DSSP (Dictionary of Secondary Structure of Proteins), que se convirtió en el estándar de oro hasta hoy (Martin *et al*, 2005). Incluso se usa en algunos de los programas de visualización de moléculas, como Rasmol y Chime (Sayle y Milner, White, 1995). Este algoritmo asigna la estructura secundaria basándose en los patrones de aparición de los enlaces de hidrógeno en las proteínas, y en la determinación del ángulo Kappa. Este ángulo se forma entre los carbonos alfa de los residuos $i-2$, i e $i+2$ del i -ésimo aminoácido. El algoritmo DSSP asigna las estructuras secundarias: Hélices (Alfa, 3-10, y Pi), Hojas Beta (paralela y antiparalela, y giros (giro Beta, doblez y estructura al azar).

STRIDE es otro algoritmo para hacer las asignaciones de estructura secundaria, desarrollado por los investigadores alemanes Frishman y Argos (1995), que se basa en los patrones de formación de enlaces de hidrógeno en las proteínas, aunque de forma diferente a DSSP. STRIDE tiene en cuenta la Energía de los enlaces y además, los ángulos de torsión Phi y Psi del esqueleto de la proteína. Estos ángulos diédricos Phi y Psi se han usado ampliamente para definir las estructuras secundarias de proteínas. El ángulo Phi está definido para el i -ésimo residuo por los átomos C_{i-1} , N_i , $C\alpha_i$ y C_i , mientras que el ángulo Psi, para el i -ésimo residuo, está definido por los átomos N_i , $C\alpha_i$, C_i y N_{i+1} .

El método SECSTR, pertenece a la familia de algoritmos tipo DSSP. Se desarrolló específicamente para mejorar la detección de las helices Pi, que constituyen aproximadamente el 4% de las proteínas (Fodje y Al-Karadaghi, 2002).

Se han desarrollado otros métodos como DEFINE (Richards, y Kundrot, 1988), que se basa sólo en carbonos Alfa ($C-\alpha$) y compara las distancias entre $C-\alpha$ del archivo pdb comparándolas con distancias ideales en estructuras secundarias. Este método también describe estructuras supersecundarias (Richards, y Kundrot, 1988).

Un algoritmo más reciente para asignar estructura secundaria, es KASKI (que significa "dos" en finlandés) propuesto por un grupo francés (Martin *et al.*, 2005), se basa en distancias entre $C-\alpha$ y los ángulos Phi y Psi alrededor de cada carbono alfa. Los autores hicieron además una interesante revisión de los diferentes métodos existentes para hacer asignación de estructura secundaria a partir de archivos pdb (Martin *et al.*, 2005).

Aunque ha habido otras propuestas metodológicas para asignar estructura secundaria a partir de las coordenadas atómicas anotadas en los archivos pdb, el mejor método sigue siendo el DSSP. Sin

embargo, la gran cantidad de propuestas es un buen indicador del hecho que la asignación sigue presentando algunas dificultades, especialmente en las zonas de transición entre estructuras secundarias regulares, y en las estructuras sin regularidad, mejor conocidas como "estructuras al azar".

Mientras que la asignación de la estructura secundaria de una secuencia de aminoácidos que está dentro de una hélice Alfa ó en hoja Beta es bastante coincidente entre los diferentes métodos de asignación de estructura secundaria. Pero es difícil asignar los residuos de aminoácidos que se encuentran al principio y al final de estas estructuras secundarias (es decir, los aminoácidos que se encuentran en los segmentos transicionales de las hélices y las hojas (Robson y Garnier, 1986).

Acerca de PFam

La base de datos de familias de proteínas (Pfam), es una gran colección de familias de proteínas, cada una representada por alineamientos múltiples de secuencia y Modelos Ocultos de Markov.

Las proteínas se componen generalmente de una o más regiones funcionales, llamadas comúnmente dominios. Diferentes combinaciones de dominios dan lugar al diverso rango de proteínas encontradas en la naturaleza. La identificación de dominios que se dan en proteínas puede permitir tener un vistazo sobre su función.

Hay dos componentes de Pfam: Pfam-A y Pfam-B. Las entradas en Pfam-A son de alta calidad; familias curadas manualmente. Es de este componente, de donde hemos tomado las proteínas agrupadas por familia. Estas entradas cubren una gran proporción de las secuencias en la base de datos. Para tener una mejor comprensión de las proteínas conocidas, se ha generado automáticamente un suplemento usando la base de datos PRODOM. Estas entradas generadas automáticamente se conocen como Pfam-B. Aunque de menor calidad, las familias de Pfam-B pueden ser útiles para identificar regiones conservadas funcionalmente cuando no se encuentran en Pfam-A.

Pfam también genera agrupamientos de alto nivel de familias relacionadas, conocidas como clanes. Un clan es una colección de entradas de Pfam-A que están relacionadas por similitud estructural ó perfiles de Modelos Ocultos de Markov. La versión actual de Pfam es la 27.0 (de Marzo de 2013, con 14831 familias). Se puede acceder a Pfam vía internet en la dirección <http://pfam.sanger.ac.uk/>.

Sobre la estimación del orden en proteínas

Entropía informacional usada como entropía estructural

Para la estimación de la entropía conformacional, aplicamos la teoría de información de Shannon (Strait y Dewey, 1996). La formulación de la entropía de información está dada por:

$$S = -\sum_{i=1}^n p_i (\ln p_i)$$

Donde i se refiere en nuestro caso a uno de los tipos de estructuras secundarias analizados (hélices, hojas, giros Beta, dobleces, random coils), donde p_i es la probabilidad de encontrar la estructura secundaria i en los resultados obtenidos; n es el número total de posibles estructuras secundarias (20) o combinaciones de las mismas según se hable de aminoácidos individuales ó tripletas de residuos; S es la entropía conformacional de cada aminoácido ó combinación de aminoácidos; esto puede aplicarse para conocer el número de estructuras secundarias en que aparece un residuo o combinación de residuos, ó también puede aplicarse para conocer el número de combinaciones de residuos que pueden dar origen a una estructura secundaria particular; es decir, se puede aplicar esta entropía conformacional, sin importar si se trata de un fragmento de secuencia ó un patrón de estructura secundaria, lo que nos permite hacer análisis en dos sentidos: desde la secuencia de residuos hacia la estructura, y viceversa.

La entropía mínima será 0, donde domina un solo tipo de estructura secundaria. La entropía máxima se obtiene cuando todos los tipos de estructura secundaria tienen igual contribución a las estructuras. (<http://14.113.239.106/~entropy/index.cgi?ln=entropy>, Shannon, 1950).

En teoría de la información, la Entropía de Shannon ó Entropía informacional es una medida de la incertidumbre asociada con una variable aleatoria. Así, cuantifica la información contenida en un mensaje, usualmente en bits ó bits/símbolo. Es la longitud mínima de un mensaje, necesaria para comunicar información (http://en.wikipedia.org/wiki/Information_entropy, Shannon, 1950). En el caso de nuestras proteínas y secuencias cortas, corresponde a la capacidad de una combinación de residuos, de formar diferentes estructuras secundarias, ó formar muy pocas estructuras, lo que nos permite inferir sobre la presencia de secuencias camaleón ó secuencias fuertemente preferentes por una estructura secundaria.

Sobre la preferencia estructural de secuencias cortas

Uno de los objetivos del presente trabajo en la formulación de un valor indicador de cuánto una secuencia corta de residuos (en nuestro caso, tripletas) puede estar determinando una estructura secundaria, a través de la dominancia con que aparece en una ó más estructuras secundarias.

En los inicios del presente trabajo, publicamos un primer acercamiento a la producción de un indicador de la preferencia estructural de una secuencia de residuos (Mejía y Tischer, 2007), que hemos llamado IPE (por Indicador de Preferencia Estructural). Ahora, hemos contrastado este valor del indicador, con el concepto de propensión de Chou y Fasman y con la Entropía conformacional, y hemos depurado el concepto, hasta encontrar que en algunos casos, permite identificar combinaciones de residuos que presentan una alta tendencia a aparecer predominantemente en una estructura secundaria y combinaciones que se acomodan a la estructura secundaria del contexto de secuencia en que aparecen en proteínas.

El indicador se calcula con base en las frecuencias de aparición de los aminoácidos ó combinaciones de los mismos, en las diferentes combinaciones de estructura secundaria (lo que hemos llamado “secuencia estructural”, en nuestro caso, “tripleta estructural”. Así, el IPE se calcula como el producto de la frecuencia más alta de una combinación de residuos, en una estructura secundaria, por la diferencia entre esta frecuencia más alta y la segunda frecuencia más alta. De esta manera, las tripletas ó combinaciones de residuos que aparecen predominantemente en una estructura secundaria, tendrán un valor de IPE cercano a 1, mientras que en el caso contrario, el IPE estará más cerca de cero.

$$\text{IPE} = F_{\text{ma}} * (F_{\text{ma}} - F_{\text{sma}})$$

F_{ma} : Frecuencia más alta. F_{sma} : Segunda frecuencia más alta. IPE: Indicador de Preferencia estructural.

Sobre Propensiones

Una vez que Christian Anfinsen reportó que había desnaturalizado la enzima Ribonucleasa, y, bajo condiciones controladas había renaturalizado la proteína, recuperando su función, fue evidente que la información para que una proteína adquiriera su forma nativa (la estructura tridimensional funcional) se encuentra en la secuencia de aminoácidos. Esto volcó buena parte de la investigación en proteínas hacia la predicción de la estructura a partir de la secuencia (Epstein et al., 1971; Anfinsen et al., 1954).

Tal vez el primer intento de predecir la estructura secundaria de una proteína a partir de la secuencia de aminoácidos, lo hicieron los investigadores Peter Chou y Gerald Fasman, de la Universidad de Brandeis, en Massachusetts, en los Estados Unidos (Chou y Fasman, 1974, a, b). Aunque trabajaron con unas pocas proteínas (las pocas cuya estructura se había resuelto en ese momento), generaron unos valores de tendencia intrínseca de cada aminoácido a formar estructuras secundarias regulares, conocido como “propensión” (para ese momento las estructuras secundarias eran: hélice Alfa, hoja plegada Beta y giros Beta).

Con base en las frecuencias de aparición de cada aminoácido en una estructura secundaria, estimaron un valor relativo de probabilidad que fue inicialmente muy utilizado (Fasman, G.D., 1987), al punto de considerarse un clásico dentro de los intentos de predicción de estructura de proteínas, aunque según muchos autores, permitía una predictibilidad de alrededor del 54%. (El valor de predictibilidad considerado aquí es el de número de aminoácidos correctamente predicho en una estructura secundaria, dividido entre el número total de aminoácidos considerado en la predicción, expresado en porcentaje). Este método de predicción tuvo muchos defensores (Gribskov *et al.*, 1986; Jameson y Wolf, 1988; Wolf *et al.*, 1988) y un buen número de detractores (Kynäjä y Valjakka, 1997; Lareo y Acevedo, 2005). Actualmente es fácil encontrar sitios de internet que ofrecen un análisis de Chou y Fasman, para cualquier secuencia de aminoácidos.

En el artículo original de Chou y Fasman, se obtuvieron valores estadísticos de propensión con base en unas cuantas proteínas. En el presente proyecto, se actualizarán estos valores para el total de proteínas incluidas en el análisis, y para familias de proteínas; también se compararán los valores de propensiones entre proteínas solubles, con proteínas de membrana; estos valores se compararán con los de Chou y Fasman, y se usarán en predicción de estructura secundaria para verificar la predictibilidad del método.

El valor de propensión se ha calculado de varias diferentes maneras. En el artículo de Chou y Fasman (1974), la propensión, para un aminoácido X en una estructura secundaria Y, se calcula así:

$$\frac{\text{\# de veces que aparece X en Y} \div \text{\# total de aminoácidos en Y}}{\text{\# total de apariciones de X} \div \text{\# total de aminoácidos}}$$

Esto permite inferir que hay algunos aminoácidos que presentan una alta tendencia a formar una estructura secundaria, y baja a formar otras, y otros aminoácidos presentan tendencias intermedias, lo que puede interpretarse como “indiferencia” a formar una estructura secundaria específica.

Acerca de estructura secundaria de proteínas

De acuerdo con el DSSP, se pueden identificar ocho estructuras secundarias regulares. En el presente trabajo, se agruparán las ocho estructuras en cinco, así:

Estructuras helicoidales

Agruparemos las hélices Alfa, 3-10 y Pi, en un grupo que llamaremos “hélices”. Estas son estructuras cuyo esqueleto de carbonos alfa y enlaces peptídicos, siguen una forma helicoidal, como una espiral, sin espacio vacío al interior de la hélice. Otra característica distintiva es la de presentar enlaces de hidrógeno estabilizadores intracatenarios, entre un aminoácido de la hélice y otro aminoácido dos, tres ó cuatro posiciones más adelante, según el tipo de hélice. Las cadenas laterales de los residuos que forman la hélice, se orientan hacia el exterior de la misma.

Hilos y hojas Beta

Las dos estructuras de hoja Beta (puentes ó hilos, y hojas Beta) que asigna el algoritmo DSSP, se agruparán en la estructura Hojas (designada aquí como E). Las estructuras de hojas presentan las siguientes características distintivas: el esqueleto de carbonos alfa y enlaces peptídicos forma una especie de zig-zag, semejando una hoja doblada como un acordeón, vista lateralmente. Esto implica que los enlaces de hidrógeno que pueden estabilizar la hoja se forman intercatenariamente; es decir, se forman con grupos al exterior de la hoja. Las cadenas laterales se orientan hacia arriba y hacia debajo de la hoja.

Giros Beta ó Giros-n

Los tres tipos de giros se considerarán tal como los asigna el DSSP: giros beta, entendidos como una especie de curva que sigue el esqueleto carbonado de la cadena peptídica, formados por tres a cuatro aminoácidos y estabilizados por un enlace de hidrógeno entre los aminoácidos 1 y 3, ó 1 y 4.

Dobléz

Es un giro con las características del giro Beta, pero sin enlaces de hidrógeno que las estabilicen, lo que le da más flexibilidad y le permite actuar como una especie de bisagra entre estructuras secundarias ó entre dominios.

Estructuras al azar

Las estructuras al azar son las que no se pueden clasificar dentro de las cuatro anteriores estructuras secundarias, y que no presentan una regularidad característica. En proteínas, el término “estructura al azar” se refiere a la conformación en que cada residuo de aminoácido en el esqueleto de una proteína (descrita por sus ángulos Phi y Psi), es independiente de las conformaciones de sus residuos vecinos. Esta es la hipótesis de par aislado de Flory (Flory’s isolated-pair hypothesis – Flory, 1953; Flory, 1969). Esta estructura empezó a definirse en 1970, cuando Tanford se refirió a las estructuras al azar en polímeros como aquellas en las que hay rotación libre alrededor de cada enlace, como se puede encontrar en muchas moléculas pequeñas (Tanford, 1970). Más recientemente, Shortle (1996), describió una estructura al azar como un estado bien definido en el que no hay interacciones entre las cadenas laterales del polímero.

Sobre estudios previos en secuencias cortas

Anishetty et al., (2002) hicieron un análisis de tripéptidos en estructuras proteicas en un conjunto de 1220 proteínas resueltas a alta resolución (≤ 2.0 Angstroms). Ellos se enfocaron en la búsqueda de rigidez de las tripletas, clasificando las tripletas en rígidas, no rígidas e intermedias, basándose en la rigidez estructural relativa entre los Carbonos Alfa y Beta de un tripéptido. Aunque las tripletas hechas de residuos hidrofóbicos tienden a ser rígidas, también encontraron algunas tripletas rígidas compuestas por aminoácidos polares. Con la noción de rigidez, buscaron detectar interacciones de cadenas laterales, pero no estructura secundaria (Anishetty et al., 2002).

Otaki et al., (2005), analizaron grupos de tres, cuatro y cinco residuos de aminoácidos en una muestra no redundante del PDB. A estas secuencias cortas les llamaron como secuencias constitutivas. Ellos encontraron que hay una distribución no aleatoria de estas secuencias cortas en las proteínas de una población.

También encontraron que las secuencias constitutivas de cinco aminoácidos aparecen en proteínas en una frecuencia mucho menor de lo esperado. Las secuencias que no aparecen en proteínas, se sintetizaron artificialmente usando el ya clásico método de Fmoc (que consiste en

sintetizar péptidos sobre una fase sólida (Merrifield, 1963; Carpino y Ham, 1972), demostrando que no presentan barreras energéticas que impidan su aparición. También se sintetizaron en un sistema de proteínas de fusión con *E. coli*, con rendimientos razonablemente altos, demostrando así que no hay un impedimento para su síntesis en sistemas biológicos. Finalmente, compararon la frecuencia de uso de estas secuencias cortas en proteínas, y la frecuencia de uso de codones del código genético. Estos resultados en conjunto sugieren que la información de las estructuras 3D, y las funciones proteicas existen en el contexto de conexiones de secuencias constituyentes cortas, y que las proteínas están compuestas por secuencias cortas constituyentes seleccionadas evolutivamente, lo que se refleja en sus diferencias en disponibilidad en la base de datos. Estos resultados deberían tener implicaciones biológicas para los estudios de estructuras proteicas (Otaki et al., 2005).

Ngan *et al.*, (2006) hicieron un análisis de tripletas de residuos usando funciones de puntaje basadas en el conocimiento para hacer predicción de estructura proteica, teniendo en cuenta el radio de curvatura formado en tripletas de residuos de aminoácidos.

Por su parte, Lareo y Acevedo (2005) realizaron un análisis de secuencias cortas de 2, 3, y 4 residuos de aminoácidos, encontrados en 16.000 hélices Alfa resueltas experimentalmente, para reevaluar el concepto de propensiones propuesto por Chou y Fasman. En su trabajo, reportan que no encontraron algunas combinaciones de las 8000 posibles combinaciones de tres residuos de aminoácidos.

1.6. METODOLOGÍA

1.6.1. Selección de la muestra de proteínas para el análisis de la frecuencia y de la estructura secundaria a partir de los archivos .pdb

El PDB es un banco de datos que actualmente tiene almacenadas unas 88000 proteínas. De ellas, se encuentran proteínas resueltas mediante varios métodos experimentales, diferentes grados de resolución, diferentes niveles de redundancia, y otros factores que requieren depuración de la información, para el análisis masivo de datos. Dado el pequeño número de veces que aparecen los residuos de Selenocisteína y Pirrolisina, decidimos trabajar sobre las proteínas que contienen los veinte aminoácidos clásicamente encontrados en proteínas. Los criterios que consideramos importantes y que hemos establecido para la uniformidad de nuestros datos, son:

Metodología de dilucidación de la estructura: Usamos proteínas cuya estructura se resolvió por la misma metodología. En el presente trabajo, usamos las proteínas resueltas por Cristalografía de Rayos X.

Grado de resolución de la estructura: seleccionamos proteínas cuya estructura 3D está solucionada a una resolución de entre 0 y 2.0 Angstroms (Å). A esta resolución se pueden distinguir las cadenas laterales, y los carbonilos del esqueleto, y también se pueden medir las distancias interatómicas con alta precisión (unos 2Å).

Nivel de redundancia aceptado: en este trabajo eliminamos la redundancia debida a que existan dos proteínas con la misma secuencia de aminoácidos, como en el caso de las proteínas multiméricas con monómeros iguales. Así que sólo analizamos proteínas cuyas secuencias eran únicas; es decir, no son exactamente iguales en su secuencia ni estructura.

Eliminación de proteínas con aminoácidos ambiguamente identificados: Se eliminaron proteínas que tenían aminoácidos ambiguos ó no resueltos, aminoácidos con asignaciones estándar para designar secuenciaciones ambiguas, del tipo "X" (para algunos autores, significa que en esa posición puede ir cualquier aminoácido, y para otros autores, significa que no se identificó con certeza si el aminoácido es Aspártico ó Asparagina); del tipo "Z" (que no se pudo resolver entre Glutámico y Glutamina) y "J" (que no se pudo distinguir entre una Leucina y una Isoleucina). Tampoco se consideraron proteínas con el aminoácido "O", Pirrolisina; ni el "U", Selenocisteína; y finalmente se eliminaron las proteínas que contuviesen el aminoácido "B", una ambigüedad entre Aspártico y Asparagina.

Eliminación de péptidos con longitudes cortas. Se eliminaron secuencias peptídicas que tuviesen menos de 20 aminoácidos, por considerarlas oligopéptidos, más que proteínas completas.

Selección de subgrupos de proteínas a partir de una muestra representativa del PDB. Además de la selección de una muestra representativa del PDB, seleccionamos grupos de proteínas con características comunes, por criterios como contenido de estructura secundaria, localización celular, Phylum al que pertenecen, organismo del cual se aislaron, tipo de enzima (en el caso de enzimas), y familia proteica según PFam. Estas agrupaciones permitieron identificar características estructurales comunes, y frecuencias de uso de aminoácidos y secuencias cortas, lo mismo que frecuencia de uso de “micromotivos” estructurales, como tripletas en estructuras secundarias particulares. Esto nos permitió comparar los resultados obtenidos entre proteínas agrupadas según un criterio (por ejemplo contenido de estructura secundaria según SCOP), con proteínas agrupadas por otro criterio (por ejemplo localización celular). Para la selección de familias, se usó la base de datos de familias de proteínas, PFam, que a partir del PDB, hace una clasificación en familias y permite entonces obtener grupos de proteínas con funciones similares y secuencias y estructuras con alto grado de similitud (Sonnhammer et al., 1997; Bateman et al., 2004; Finn, 2005).

1.6.2. Asignación de estructura secundaria a partir de las coordenadas cartesianas de las posiciones de los átomos de cada proteína

Para asignar la estructura secundaria a una secuencia de proteína del PDB, se usó el algoritmo DSSP, que es el estándar de oro que usa el PDB para la asignación de estructura secundaria a partir de las coordenadas cartesianas de localizaciones de los átomos de las proteínas. Aunque el DSSP asigna 8 tipos de estructuras secundarias, en el presente proyecto trabajamos con 5 estructuras secundarias, uniendo en una sola los tres tipos de hélices del DSSP; los dos tipos de estructuras de hoja Beta en una sola, y manteniendo los tres tipos de giro que el DSSP identifica.

Se tuvieron en cuenta las zonas de transición entre una estructura secundaria y otra, y las secuencias cortas cuya asignación en estructura secundaria se puede considerar como de difícil interpretación biológica y estructural, como las tripletas en que un solo aminoácido aparece en hoja beta, entre aminoácidos de hélice alfa (de la forma "HBH"), dado que no tiene sentido hablar de una hoja Beta de una longitud de un solo residuo de aminoácido.

1.6.3. Determinación de las frecuencias de aparición de aminoácidos individuales y tripletas, y relación entre estas frecuencias y la estructura secundaria

Se encontraron secuencias tripletas de residuos de aminoácidos con una muy alta frecuencia de aparición, y otras con una muy baja. Se calcularon las frecuencias en la muestra representativa del PDB, y en las proteínas del PDB agrupadas con los criterios establecidos arriba. Se verificó si las frecuencias de la muestra del PDB, se conservan dentro de las proteínas agrupadas. Se identificaron los grupos de proteínas con frecuencias diferentes a la totalidad de las proteínas, representadas en la muestra representativa. La clasificación en grupos de proteínas permitió identificar tripletas que tienen una relación con el criterio de agrupación, como localización celular, contenido de estructura secundaria, tipo de enzima, etc. Se verificó que algunas combinaciones de aminoácidos en tripletas, muestran una tendencia a cambiar según diferentes criterios de agrupación de proteínas (por ejemplo, que el Triptófano es más frecuentemente encontrado en proteínas bacterianas que en proteínas de organismos superiores), ó la localización celular (porque el ambiente afecta el plegamiento y posiblemente, se utilizan secuencias distintas en diferentes ambientes).

1.6.3. Análisis de la propensión de secuencias cortas de aminoácidos relacionadas con las secuencias estructurales.

Para analizar la propensión, se determinaron las frecuencias de aparición de las combinaciones de tres residuos de aminoácidos (en adelante, tripletas de aminoácidos), en las estructuras secundarias de las proteínas seleccionadas, y con base en ellas, se calculó un valor de propensión de cada secuencia corta de residuos, generalizando lo propuesto por Chou y Fasman (1978) para aminoácidos individuales, y extendiéndolo a diferentes tipos de giros, completando cinco estructuras secundarias, y a las zonas de transición entre una estructura secundaria y otra. Se calculó un Indicador de Preferencia Estructural, como se calculó anteriormente para un conjunto menor de proteínas (Mejía y Tischer, 2007). Se calculó el valor de entropía estructural, usando la fórmula de cálculo de entropía informacional de Shannon (1950). Se compararon estos diferentes indicadores usando modelos estadísticos como el análisis de correlación.

1.6.4. Identificación y caracterización de zonas de mayor ó menor estabilidad respecto de la estructura secundaria.

Los indicadores mencionados en el punto anterior se usaron para definir las zonas que tienden a permanecer estables en una estructura secundaria específica y las zonas que no presentan esa estabilidad. De esta manera, se compararon proteínas cuya estructura presenta alta estabilidad, con proteínas cuya estructura secundaria puede cambiar, y se verificó que mediante nuestro

indicador, podemos identificar zonas que tienden a permanecer en una estructura secundaria, ó zonas que pueden cambiar su conformación de estructura secundaria.

1.6.4. Sobre las tripletas, la generación de alfabetos reducidos, y las agrupaciones de las estructuras secundarias.

En el presente trabajo, además de los análisis sobre los 20 aminoácidos, las 8000 combinaciones de tres residuos, y las 125 estructuras secundarias resultantes de las combinaciones de 5 estructuras secundarias en grupos de 3 residuos, hicimos también análisis sobre alfabetos reducidos, de la siguiente manera:

Los veinte aminoácidos se agruparon así:

Los aminoácidos no polares (hidrofóbicos), Alanina, Fenilalanina, Leucina, Isoleucina, Valina y Metionina, se agruparon en una sola categoría, que designamos con la letra “U”.

Los aminoácidos cuya cadena lateral es básica y se ionizan cargándose positivamente, Histidina, Leucina, Arginina, se agruparon en la categoría “B”.

Los aminoácidos cuya cadena lateral es un hidroxilo, que además pueden participar en modificaciones postraduccionales, e interacción con moléculas de agua, Serina, Treonina y Tirosina, que se agruparon en “O”.

Los aminoácidos con cadena lateral ácida, que al ionizarse quedan cargados negativamente, ácido Aspártico y ácido Glutámico, agrupados en la letra “X”.

Los aminoácidos cuya cadena lateral es un grupo amida: Asparragina y Glutamina, que se agruparon en la categoría “Z”.

Algunos aminoácidos no se agruparon debido a que consideramos que tienen características únicas y poco sustituibles en proteínas, y quedaron cada uno en su propia categoría: Glicina (“G”), por tener la cadena lateral más pequeña, que permite mucho acercamiento entre zonas de una proteína. El Triptófano (“W”), por ser el aminoácido más grande, por su carácter no polar, por estar codificado por un solo residuo y por tener una frecuencia de aparición muy baja en proteínas. La Cisteína (“C”), por ser el único aminoácido que puede formar puentes de disulfuro, contribuyendo con este enlace covalente, a la estabilidad de la estructura tridimensional de muchas proteínas. El aminoácido Prolina (“P”), por su carácter como iminoácido y por formar enlaces peptídicos con una torsión debida a su grupo imino secundario, y por tanto introducir torsiones en las estructuras secundarias de las proteínas.

Con esto, de un alfabeto de 20 aminoácidos, pasamos a uno de 9 letras, con lo que algunos análisis mostraron tendencias que no se notaban fácilmente al considerar los 20 residuos.

Esta reducción del alfabeto de aminoácidos, nos permitió hacer un análisis adicional y pasar de 8000 tripletas ($20^3 = 8000 = 20$ aminoácidos en combinaciones de 3), a 729 tripletas ($9^3 = 729$ agrupaciones de aminoácidos en combinaciones de 3).

También redujimos el análisis de 125 posibles estructuras secundarias al combinar 3 residuos de aminoácidos (por $5^3 = 125$ estructuras secundarias analizadas, en combinaciones de 3 residuos), a 26 tripletas estructurales, al agrupar las transiciones de una estructura a otra, y separar las tripletas cuya asignación presenta una difícil interpretación bioquímica (de la forma XYX , ó XYZ , pues no tiene sentido hablar de una estructura secundaria de una longitud de 1 residuo. Ej una hélice de 1 residuo).

1.7. BIBLIOGRAFÍA - DESCRIPCIÓN DEL TRABAJO

Altschul, S.F. (1991). **Amino acid substitutions matrices from an information theoretic perspective.** J. Mol. Biol. 219, 555-665.

Anfinsen, C.B., Redfield, R.R., Choate, W.I., Page, J., y Carroll, W.R. (1954). **Studies on the gross structure, cross-linkages, and terminal sequences in Ribonuclease.**" Journal of Biological Chemistry 207, 1: 201-210.

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS (January 2004). **"UniProt: the Universal Protein knowledgebase"**. Nucleic Acids Res. 32 (Database issue): D115–9.

Anishetty, S., Pennathur, G. y Annishetty, R. (2002). **Tripeptide analysis of protein structures.** BMC Structural Biology. 2:9-16.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L., Studholme, D.J., Yeats C. y Eddy S.R. (2004). **The Pfam Protein Families Database.** Nucleic Acids Research, 32:D138-D141.

Bateman, A. Birney, E., Cerruti, L. Durbin, R. Etwiller, L. Eddy, S.R. Griffiths-Jones, S. Howe, K.L. Marshall, M. y Sonnhammer, E.L. (2002). **The Pfam Protein Families Database.** Nucleic Acids Research 30(1):276-280.

Bateman, A. Birney, E., Durbin, R., Eddy, S.R., Howe K.L. y Sonnhammer E.L. (2000). **The Pfam Protein Families Database:** Nucleic Acids Research 28:263-266.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R. Finn R.D. y Sonnhammer, E.L.L. (1999). **Pfam 3.1: 1313 multiple alignments match the majority of proteins.** Nucleic Acids Research 27:260-262.

Berman, H., Westbrook, J., Feng, Z., Guilliand, G., Bhat, T.N., Weissing, H., Shindyalov, I.N., y Bourne, P.E. (2000). **The Protein Data Bank.** Nucleic Acids Research. 28 (1) : 235-242.

Berman, H.M., Henrick, K. y Nakamura, H. (2003). **Announcing the worldwide Protein Data Bank.** Nat Struct Biol, 10: 98.

Carpino, L. A. y Han, G. Y. (1972). **The 9-fluorenylmethoxycarbonyl aminoprotecting group.** J. Org. Chem. 37, 3404–3409.

Chou, P.Y., y Fasman, G.D. (1974a). **Prediction of protein conformation.** Biochemistry—USA 13:222-45.

Chou, P.Y., y Fasman, G.D.(1974b). **Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins.** Biochemistry—USA 13:21-35. 1974.

Chou, PY y Fasman, GD. (1974c). **Prediction of protein conformation**. Biochemistry, 13, 222-245.
El artículo clásico de Chou y Fasman.

Chou, PY. (1989) En Fasman, GD. (ed.), **Prediction of protein structure and the principles of protein conformation : Prediction of protein structural classes from amino acid compositions**. New York, pp. 549-586.

Chou PY, Fasman GD. (1977). **Prediction of protein conformation: secondary structure**. Journal of Molecular Biology., 115, 135-175.

Coin, L., Bateman A. and R. Durbin. (2003). **Enhanced protein domain discovery by using language modeling techniques from speech recognition**. Proc. Natl. Acad. Sci. USA. 100(8):4516-4520.

Chuan C.H. (2012). **Proteomics in Clinical Studies and its Applications**. Single Cell Biology: 1-2.

Djikaev, Y. S. y Ruckenstein, E. **Model for the Nucleation Mechanism of Protein Holding**. Journal of Physical Chemistry B, 111 (4), 886 -897, 2007.

Dill, K. A., Bromberg, S., Yue, K. Z., Fiebig, K. M. Yee, D. P., Thomas, P. D. y Chan, H. S. (1995). **Principles of protein folding - A perspective from simple exact models**. Protein Science., 4, 561-602.

Epstein, H.F., Schechter, A.N., Chen, R.F., y Anfinsen, C.B. (1971). **Folding of staphylococcal nuclease: Kinetic studies of two processes in acid renaturation**. Journal of Molecular Biology, Volumen 60 (3): 499-508.

Fasman, G.D., 1989. Libro: **Prediction of Protein Structure and the Principles of Protein Conformation**. Pag 221. (ISBN 0306431319) New York (Plenum).

Fasman, G.D., (1987). Biochemistry, sección : "This week's citation classic". 30 : 32 : 21 (Agosto 10).

Finn R.D. (2005). **Pfam: the protein families database**. (eds M.J. Dunn, L.B. Jorde, P.F.R. Little, S. Subramaniam) en: **Genetics, Genomics, Proteomics and Bioinformatics, Section 6: Protein Families** ISBN 978-0-470-84974-3 2005.

Finn, J. Mistry, B. Schuster-Böckler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S.R. Eddy, E.L.L. Sonnhammer and A. Bateman (2006). **Pfam: clans, web tools and services**. Nucleic Acids Research Database Issue 34:D247-D51.

Finn, R.D. Bateman, A. y Griffiths-Jones, S. (2008) **Identifying protein domains with the Pfam database**. Current protocols in bioinformatics ISBN 978-0-471-25093-7. Chapter 2:Unit 2.5.

Flory, P.J. (1953). **Principles of polymer chemistry**. Ithaca, NY: Cornell University Press.

Flory, P.J. (1969). **Statistical mechanics of chain molecules**. New York: Wiley.

Fodje MN, Al-Karadaghi S. (2002). **Occurrence, conformational features and amino acid propensities for the pi-helix**. Protein Engineering, 15(5):353-358.

Frishman D, Argos P. (1995). **Knowledge-based protein secondary structure assignment**. Proteins, 23(4):566-579. <http://webclu.bio.wzw.tum.de/stride/>.

Garnier, J., Osguthorpe, D.J. y Robson, B. (1978). **Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins**. Journal of Molecular Biology. 120: 97–120.

Giacovazzo, C.; Monsaco, H. L.; Artioli, G.; Viterbo, D.; Ferraris, G.; Gilli, G.; Zanotti, G.; Catti, M. (2002). **Fundamentals of Crystallography**; 2° ed. Giacobazzo, C., Ed.; Oxford : Oxford University Press.

Gong, H y Rose, G.D. (2005). **Does secondary structure determine tertiary structure in proteins?** en “Proteins: Structure, Function, and Bioinformatics”, Volume 61 (2), 338 – 343.

Gribskov, M, Burgess, R.R., J. Devereux. (1986). **PEPLOT, a protein secondary structure analysis program for the UWGCG sequence analysis software package**, Nucl Ac Research 14 : 327-334.

Henikoff, S. and Henikoff, J. (1992). **Amino acid substitution matrices from protein blocks**. Proc. Natl. Acad. Sci. USA. 89 (biochemistry): 10915 – 10919.

Henrick, K., Feng, Z., Bluhm, W.F., Dimitropoulos, D., Doreleijers, J.F., Dutta, S., Flippen-Anderson, J.L., Lonides, J., Kamada, Ch., Krissinel, E., Lawson, C.L., Markley, J.L., Nakamura, H., Newman, R., Shimizu, Y., Swaminathan, J., Velankar, S., Ory, J., Ulrich, E.L., Vraken, W., Westbrook, J., Yamashita, R., Yang, H., Young, J., Yousufuddin, M. y Berman, H.M. (2007). **Remediation of the protein data bank archive**. Nucleic Acids Research. Dec 2007: 1-8.

ICGEB Symposium: “The Biology in the Post-genomic era”. Trieste, Italy, 2001.

Jameson, B.A. y H. Wolf. (1988). **The antigenic index: a novel algorithm for predicting antigenic determinants**. Computer Applications in the Biosciences 4 : 181-186.

Kabsch W, Sander C. (1983). **Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features**. Biopolymers, 22(12):2577-637.

Kendrew J, Bodo G, Dintzis H, Parrish R, Wyckoff H, Phillips D. (1958). **A three-dimensional model of the myoglobin molecule obtained by x-ray analysis**. Nature 181 (4610): 662–6.

Kyngäs, J. y Valjakka, J. (1997). **Why the Chou-Fasman parameters are not reliable in protein secondary structure prediction**. The Pennsylvania State University CiteSeer Archives. Referencia completa, resto de la referencia. Volumen, Revista, páginas, año, etc.

Lareo, L.R. y Acevedo, O.E. (2005). **Amino acid propensities revisited**. OMICS, A Journal of Integrative Biology. 9, 4: 391-399.

Leszczynski, J.F., Rose, G.D. (1986). **Loops in globular proteins: A novel category of protein secondary structure**. Science 234:849–855, 1986.

Levitt M, Greer J. (1977). **Automatic identification of secondary structure in globular proteins**. Journal of Molecular Biology. 114(2):181-239.

Lobley A, Swindells MB, Orengo CA, y Jones DT. (2007). **Inferring function using patterns of native disorder in proteins**. PLoS Comput Biol 3(8): e162.

Martin, J., Letellier, G., Marin, A., Taly, J.F., de Brevern, A., y Gibrat, J.F. (2005). **Protein secondary structure assignment revisited: a detailed analysis of different assignment methods**. BMC Structural Biology, 5:17.

Mejía, D.F, Tischer, I. (2007). **Design of a new indicator of structural preference for residue triplets into proteins**. Capítulo del libro: “Proceedings, International Seminar on Genomics, Proteomics, Bioinformatics and Systems Biology” (ISBN 978-958-9451-21-2), pages 123-130. Popayán, Colombia.

Merrifield, R.B. (1963). **Solid Phase Peptide Synthesis. I. The Synthesis of a Tetrapeptide**. Journal of the American Chemical Society 85: 2149.

Meyer, E.F. (1997). **The first years of the Protein Data Bank**. Protein Science. 6: 1591-1597.

Mistry, J. y Finn R.D. **Pfam: a domain-centric method for analysing proteins and proteomes**. (2007). Methods in Molecular Biology - Comparative Genomics. 396:43-58.

Mossos, N., Mejia-Carmona, D.F., Tischer, I. (2014). **FS-Tree: Sequential Association Rules and First Applications to Protein Secondary Structure Analysis**. Advances in Computational Biology - Advances in Intelligent Systems and Computing. 232: 189-198.

Muirhead H, Perutz M. (1963). **Structure of hemoglobin - a three-dimensional fourier synthesis of reduced human hemoglobin at 5.5 Å resolution**. Nature 199 (4894):633–638.

Myers, J.K., y Oas, T.G. (2001). **Preorganized secondary structure as an important determinant of fast protein folding**. Nature Structural Biology 8, 552 – 558.

Nakayama, K., Tachikawa, K., Majima, T. (2008). **Protein Recording Material: Photorecord/Erasable Protein Array Using a UV-Eliminative Linker**. Langmuir – American Chemical Society. 24: 1625-1628.

Ngan, S-C., Inouye, M.T., y Samudrala, R. (2006). **A knowledge-based scoring function based on residue triplets for protein structure prediction**. Protein Engineering, Design & Selection vol. 19 no. 5 pp. 187–193.

Otaki, J.M., Ienaka, S., Gotoh, T. y Yamamoto, H. (2005). **Availability of short amino acid sequences in proteins**. Protein Science. 14:617-625.

PDB Annual Report 2007, Enero (2008).

Protein Data Bank. (1971). **Protein Data Bank**. Nature New Biol. 233:223.

Protein Data Bank contents guide: Atomic coordinate entry format description, Version 3.1, July 19, 2007. Liberado en Agosto 1 de 2007. <http://www.wwpdb.org/documentation/format3.1-20070719.pdf>,

Richards FM, Kundrot CE. (1988). **Identification of structural motifs from protein coordinate data: secondary structure and first level supersecondary structure**. Proteins, 3(2):71-84.

Risso, V., Gaviria, JA., **Mejia-Carmona, DF.**, Gaucher, E., Sanchez-Ruiz, JM. (2013). **Precambrian Antibiotic Resistance**. Federation of European Biochemical Societies – Congress 2013, July, 2013.

Risso, V., Gaviria, JA., **Mejia-Carmona, DF.**, Gaucher, E., Sanchez-Ruiz, JM. (2013). **Hyperstability and enhanced promiscuity in resurrected Precambrian β -Lactamases**. Journal of the American Chemical Society. 135(8):2899-902.

Robson B, Garnier J. (1986). **Introduction to Proteins and Protein Engineering**. Amsterdam: Elsevier Press.

Sayle RA, Milner-White EJ: (1995). **RASMOL: biomolecular graphics for all**. Trends in Biochemical Science, 20(9):374.

Shannon, Claude E. (1950): **Prediction and entropy of printed English**, The Bell System Technical Journal, 30:50-64.

Shortle, D. **The denatured state (the other half of the folding equation) and its role in protein stability**. The FASEB journal. (1996), 10, 1, 27-34.

Sonnhammer, E.L.L., Eddy S.R. y Durbin R. (1997). **Pfam: a comprehensive database of protein families based on seed alignments**. Proteins. 28:405-420.

Sonnhammer, L.L., Eddy, S.R., Birney, E., Bateman, A. y Durbin, R. **Pfam: multiple sequence alignments and HMM-profiles of protein domains**. Nucleic Acids Research (1998) 26:320-322.

Strait, B.J. y Dewey, T.G. (1996). **The Shannon information entropy of protein sequences**. Biophysical Journal. 71(1): 148–155.

Tanford, C. (1970). **Protein denaturation**. Advances in protein chemistry, 24:1-95.

Wolf, H., Modrow, S., Motz, M., Jameson, B.A., Hermann G. y B. Fortsch. (1988). **An integrated family of amino acid sequence analysis programs**. Computer Applications in the Biosciences 4: 187-191.

CAPÍTULO 2. SELECCIÓN DE LAS MUESTRAS

En este capítulo se definen los conjuntos de datos usados en esta investigación. Se define un conjunto global de proteínas del cual se extraen subconjuntos que permiten estudiar diferentes aspectos que pueden afectar la composición respecto a los aminoácidos y las estructuras secundarias consideradas. Se verifica que el conjunto global de proteínas seleccionadas puede considerarse una muestra representativa, al compararlo con conjuntos más grandes que representan el universo proteico.

2.1. La muestra representativa del PDB

El PDB tiene almacenadas actualmente (a Noviembre de 2013), unas 88.000 proteínas. De ellas, se encuentran proteínas resueltas mediante diferentes métodos experimentales, con diferentes niveles de resolución, diferentes niveles de redundancia, y otros factores que requieren depuración de la información.

Las muestras seleccionadas cumplen con los siguientes criterios de búsqueda: seleccionamos moléculas de proteína sin ácidos nucleicos como ligandos; estas proteínas debían estar resueltas por Cristalografía de Rayos X, a una resolución de 2.0 Angstroms (Å) o mejor, resueltas a niveles de detalle de entre 0 y 2.0 Å; se eliminaron las proteínas iguales, bien sea porque diferentes autores podrían haber registrado la misma proteína en el PDB, ó porque algunas proteínas de diferentes organismos, pudiesen tener exactamente la misma secuencia y estructura; se eliminaron las proteínas que contienen aminoácidos clasificados como “B”, “J”, “O”, “U”, “X” y “Z” (“indeterminado entre Asparagina (N) y Glutamina (Q)”), que corresponden a aminoácidos que no pudieron ser claramente determinados, o a Pirrolisina o Selenocisteína, que decidimos no incluir en el estudio, por razones discutidas en la metodología. Finalmente, se eliminaron los péptidos pequeños, que no pueden considerarse una proteína completa, por lo que escogimos tamaños mínimos de 21 residuos de aminoácidos (Tabla 2.1.1.).

Tabla 2.1.1. Criterios tenidos en cuenta para la selección de las proteínas del presente estudio.

Criterio considerado	# de moléculas cumpliendo el criterio
Proteína, sin DNA ó RNA como ligando.	80681
Moléculas resueltas por Rayos X.	76732
Resolución entre 0 y 2 Angstroms.	36695 (8196 ligandos)
Se eliminaron las proteínas idénticas al 100%.	
Proteínas con errores en la asignación de DSSP.	59
Proteínas con aminoácidos ambiguos (B,J,O,U,X,Z).	3
Total de proteínas seleccionadas:	19308

El conjunto seleccionado bajo estos criterios, que llamamos **PDBGlobal**, puede considerarse como una muestra representativa del PDB, como se aclarará a lo largo de este capítulo. Dentro de este

conjunto se seleccionaron subconjuntos que incluyen subconjuntos de las familias de proteínas con características comunes en términos de su estructura, función, localización celular, u organismo del que provienen, lo cual permite comparar los resultados obtenidos, entre diferentes propiedades de las proteínas seleccionadas (ej. Permite comparar la secuencia y estructura secundaria de Enzimas deshidrogenasas, con proteínas transportadoras de oxígeno, ó con proteínas de la matriz extracelular), con el fin de identificar particularidades respecto de la relación secuencia y estructura secundarias, no evidentes a simple vista.

Nuestra muestra representativa del PDB, que hemos llamado **PDBGlobal**, incluyó 19.308 proteínas, con un total de 8'862.961 de aminoácidos, que sumaron un total de 8'784.695 tripletas de residuos.

La muestra **PDBXRay** (Todas las proteínas del PDB, resueltas por Cristalografía de Rayos X), abarcó 75.215 proteínas, con un total de 22'959.577 de aminoácidos, que sumaron un total de 22'782.781 tripletas de residuos.

El universo proteico almacenado en el PDB hasta la fecha de la última actualización de datos para este trabajo (Junio 23 de 2013), que hemos llamado **PDBUniverso**, abarcó 82.893 proteínas, con un total de 38'475.153 residuos de aminoácidos, que sumaron un total de 37.939.740 tripletas de residuos.

Debido a que el número de proteínas de nuestra muestra PDBGlobal puede considerarse relativamente bajo con respecto a la totalidad del PDB, para se pueda considerar representativo, comparamos las frecuencias de los aminoácidos individuales de estas proteínas, con las de una muestra de todas las proteínas del PDB resueltas por Cristalografía de Rayos X (PDBXRay, con 75.215 proteínas), y con la totalidad de las proteínas asentadas en el PDB hasta Junio 23 de 2013 (PDBUniverso, 82.893 proteínas). La comparación entre las frecuencias de las tres muestras, se presenta más adelante en este capítulo.

En el estimado de las frecuencias de los aminoácidos individuales, sin asociarlas con la estructura secundaria, también hicimos una comparación con el banco de datos de secuencias proteicas, UniProtKB/Swiss-Prot, que en su versión del 6 de Marzo de 2013 (<http://web.expasy.org/docs/relnotes/relstat.html>), contenía 539.616 secuencias proteicas, con 191.569.459 aminoácidos (Tabla 2.1.1.).

2.2. La muestra de proteínas agrupadas por su contenido de estructura secundaria según SCOP.

Para la selección de proteínas por su contenido de estructura secundaria, se recurrió a la base de datos SCOP (Secondary Structure of Proteins). La clasificación SCOP (por Structural Classification of

Proteins), propuesta en el Laboratorio de Biología Molecular y Centro para la Ingeniería de Proteínas de la Universidad de Cambridge (Murzin et al., 1995) y actualizada por los mismos autores (Lo Conte et al., 2002; Andreeva, 2004; Andreeva, 2007), es considerada una de las clasificaciones estructurales más completas de los dominios de las proteínas. SCOP es usada por el PDB para la organización de las proteínas por su contenido de estructura secundaria, así que, en el PDB se pueden encontrar las proteínas agrupadas en las clases que corresponden a nuestra selección.

SCOP, la base de datos de las clasificaciones estructurales de proteínas, en su versión versión 1.75, cuenta con 38.221 proteínas en el PDB (a 23 Feb de 2009), 110.800 dominios, excluyendo ácidos nucleicos y modelos teóricos (tomado de <http://scop.mrc-lmb.cam.ac.uk/scop/count.html#scop-1.75> y consultado el 15 de Septiembre, 2013). Esta muestra de las proteínas del PDB, y de SCOP, abarca proteínas de diferentes familias y superfamilias, y se presenta en la tabla 2.2.1.

Tabla 2.2.1. Estadísticas de SCOP (Structural Classification of Proteins), versión 1.75.

Clase	# de plegamientos	# de Superfamilias	# de Familias
All Alpha	284	507	871
All Beta	174	354	742
Alpha and Beta (a/b)	147	244	803
Alpha and beta (a+b)	376	552	1055
Multi-domain	66	66	89
Membrane and cell surface	58	110	123
Small proteins	90	129	219
Total	1195	1962	3902

De acuerdo con la clasificación SCOP, y cumpliendo con las premisas para la selección de los datos para nuestro estudio, la muestra de proteínas agrupadas por su composición de estructura secundaria, y descritas en la tabla 2.2.2., son: **Proteínas Todo Alfa** (que en algunas partes de este documento, hemos llamado indistintamente “Todo Alfa” ó “AllAlpha”): 1.399, **Todo Beta** (ó “AllBeta”): 1.771, **Alfa+Beta** (ó “Alpha+Beta”): 2.066, **Alfa/Beta** (ó “Alpha-Beta”): 2.577, **Proteínas multidominio** (o “Multidomain”): 175, **Proteínas pequeñas** (o “Small”): 217, **Proteínas “Coiled coil”**: 97, y proteínas clasificadas como **“Otras”** (“others”): 113. Las clasificaciones de las proteínas en SCOP, se basaron en los criterios de la tabla siguiente (tabla 2.2.2., considerando que la unidad para clasificar en SCOP, es el dominio proteico:

Tabla 2.2.2. Criterios para la clasificación de proteínas por su contenido de estructura secundaria según SCOP, y número de proteínas seleccionadas.

Clasificación de proteínas según SCOP	# de proteínas	Criterio para la clasificación
Todo-Alfa	1.399	Proteínas cuyo contenido en hélices Alfa es dominante y

		contiene poco o nada de hilos u hojas Beta.
Todo-Beta	1.771	Proteínas cuyo contenido en hilos u hojas Beta es dominante y contiene poco o nada de hélices Alfa.
Alfa y Beta (α/β)	2.066	Contienen principalmente hojas Beta <u>paralelas</u> (Unidades tipo hilo β - hélice α - hilo β). Aquí se designan como Alfa-Beta.
Alfa y Beta ($\alpha+\beta$)	2.577	Contienen principalmente hojas Beta <u>antiparalelas</u> (regiones segregadas de hélice α y hoja β)
Multidominio (α y β)	175	Plegamientos de dos ó más dominios de diferentes clases.
Pequeñas ("small")	217	Contienen un solo dominio y están usualmente dominadas por un ligando metálico, un grupo hemo y/o puentes de disulfuro.
"coiled coil"	97	No se considera una verdadera clase de proteínas. Sin embargo, es un motivo estructural que se da cuando de dos a siete hélices alfa están enrolladas juntas, como las hebras de una cuerda.
Otras proteínas	113	En el banco de datos del PDB, las proteínas que no se incluyen en las anteriores clases, se consideran "otras".

2.3. La muestra de proteínas agrupadas por su localización celular

El PDB ofrece la posibilidad de seleccionar sus proteínas por diferentes criterios. Entre ellos, tomamos ventaja de la posibilidad de seleccionar proteínas por su localización celular. Las proteínas seleccionadas según este criterio, se presentan en la tabla 2.3.1.

Tabla 2.3.1. Proteínas seleccionadas para agruparlas por localización celular.

Localización celular	# de proteínas
Matriz extracelular	57
Matriz mitocondrial	98
Complejo de Golgi	160
Membranales	917
Otras organelas	694
Citoplasmática	3.859
Citoesqueleto	237
Membrana interna mitocondrial	51

2.4. La muestra de proteínas agrupadas por el organismo del que provienen

Para la selección de estos grupos de proteínas, tuvimos en cuenta representantes de eucariotes, procariotes y archaea. Además, seleccionamos los organismos que en el PDB tienen un mayor

número de proteínas resuelto según los criterios con que seleccionamos la muestra PDBGlobal. También procuramos seleccionar organismos que viviesen en ambientes extremos, como *Thermus thermophilus* una bacteria termófila Gram negativa, aislada de fumarolas hidrotermales, cuya temperatura óptima de crecimiento es de unos 65°C (Oshima e Imahori, 1974) y *Thermotoga maritima*, una bacteria termófila Gram negativa muy antigua encontrada en termales, con una temperatura óptima de crecimiento de 80°C, con el fin de estudiar las proteínas que presenten características interesantes, como la termoestabilidad. Las proteínas agrupadas por organismo y seleccionadas para el presente estudio, fueron:

Tabla 2.4.1. Proteínas seleccionadas y agrupadas por el organismo del que se provienen.

Organismo	# de proteínas
<i>Thermus thermophilus</i>	233
<i>Thermotoga marina</i>	263
<i>Mycobacterium tuberculosis</i>	124
<i>Pseudomonas aeruginosa</i>	192
<i>Bacillus subtilis</i>	336
<i>Escherichia coli</i>	818
<i>Saccaromyces cerevisiae</i>	492
<i>Drosophila melanogaster</i>	142
<i>Arabidopsis thaliana</i>	185
<i>Mus musculus</i>	658
<i>Rattus norvegicus</i>	281
<i>Bos taurus</i>	123
<i>Homo sapiens</i>	3.325

En esta submuestra, hay representantes de los procariotes (bacterias patogénicas, no patogénicas y organismos termófilos), levaduras, plantas, insectos y mamíferos. No se consideraron otros organismos, ya que el PDB contiene muy pocas proteínas de estos (en algunos casos, menos de 10).

2.5. La muestra de proteínas agrupadas por el Phylum al que pertenece el organismo del que provienen

Una de las agrupaciones de proteínas, fue la del Phylum del organismo al que pertenecen las proteínas. Una vez más, sacamos ventaja de las posibilidades de clasificación de los datos, que ofrece el PDB. Las proteínas seleccionadas según el phylum, se presentan en la Tabla 2.5.1.

Tabla 2.5.1. Proteínas seleccionadas y agrupadas por el Phylum al que pertenece el organismo del que se provienen.

Phylum	# de proteínas
Virus	756
Archaea	1.120
Procariotes	10.361
Eucariotes	7.824

2.6. La muestra de proteínas agrupadas por el tipo de enzima según la nomenclatura enzimática

Las proteínas con actividad catalítica (ya una primera clasificación), se subclasificaron según el tipo de enzima, de acuerdo con el Comité de Nomenclatura de la Unión Internacional de Bioquímica y Biología Molecular (NC-IUBMB, por su sigla en inglés “Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB)”. Esta nomenclatura se conoce como nomenclatura de la Comisión de Enzimas (Enzyme Commission, EC.) (Webb, 1992; Moss, 1992). Las enzimas agrupadas según la nomenclatura a la que pertenece la reacción química realizada por la enzima, se presentan en la siguiente tabla (Tabla 2.6.1.)

Tabla 2.6.1. Proteínas con actividad catalítica, clasificadas por tipo de enzima, de acuerdo con el “Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB)”:

Tipo de enzima EC	# de proteínas
EC1. Oxidoreductasas	1.730
EC2. Transferasas	2.628
EC3. Hidrolasas	2.975
EC4. Liasas	780
EC5. Isomerasas	571
EC6. Ligasas	477

2.7. La muestra de proteínas agrupadas por la familia a la que pertenecen, según Pfam

Para la selección de familias, se usó la base de datos de familias de proteínas, Pfam, que a partir del PDB, hace una clasificación en familias y permite entonces obtener grupos de proteínas con funciones similares y secuencias y estructuras con alto grado de similitud (Sonnhammer et al., 1997; Bateman et al., 2004; Finn, 2005). De estas familias de proteínas, se seleccionaron grupos de proteínas que presentan un alto contenido de hilos y hojas Beta, y que además presentan una relativamente alta estabilidad, como la del grupo de la Proteína Verde Fluorescente (GFP); otras con un alto contenido de hélices, como las Globinas; otras conocidas por su alta estabilidad térmica, como las proteínas hiperestables CutA1; otras por su alta estabilidad a bajas temperaturas, como las proteínas Anticongelantes (AFP por AntiFreezing Proteins); proteínas muy

conocidas y caracterizadas como Actina; proteínas con un alto contenido de Cisteínas, como las proteínas CRISP (por Cysteine-rich secretory proteins); proteínas de interés de nuestros proyectos en ingeniería de proteínas, como los Factores de Crecimiento transformante tipo Beta (TGFB), específicamente con las proteínas morfogenéticas de hueso (TGFB-BMP), Fosfolipasas A2 de venenos de serpientes (PLA2), y Beta-Lactamasas. Aunque algunas proteínas presentan poco número de moléculas caracterizadas en el PDB, hemos considerado que su particularidad estructural (como la hiperestabilidad, ó el contenido de Cisteínas), hizo que las incluyéramos en el presente estudio. Los resultados de la selección de estas proteínas, se encuentran en la tabla 2.7.1., a continuación.

Tabla 2.7.1. Proteínas seleccionadas y agrupadas por familia proteica, según PFam.

Familia de proteínas	# de proteínas
Actina	75
AFP	79
Beta Lactamasas	178
CRISP	11
GFP	123
Globina	178
Hyperstable CutA1	27
PLA2	199
Prión	80
TGFB-BMP	200

2.8. Verificación de la representatividad de la muestra PDBGlobal

La muestra PDBGlobal, se comparó en las frecuencias de sus proteínas, con las frecuencias encontradas en la totalidad de las proteínas resueltas en el PDB, mediante Cristalografía de Rayos X (la muestra PDBXRay), la totalidad de las proteínas del PDB al momento de la última actualización del presente estudio (PDBUniverso), y una base de datos con un número mucho mayor de proteínas almacenadas, al menos en su secuencia: UniProt/SwissProt (Apweiler et al, 2004).

2.8.1. Comparación de las frecuencias de residuos entre PDBGlobal, PDBXRay, PDBUniverso y UniProt/SwissProt

Se analizan las frecuencias de aparición de los veinte aminoácidos estándar, contados individualmente. Se calcularon las frecuencias de los veinte aminoácidos estándar encontrados las

proteínas. Para ello, se usó la muestra PDBGlobal (19.308 proteínas, 35782 secuencias proteicas, 2'033.850 aminoácidos) y se comparó con la totalidad de las proteínas resueltas por Cristalografía de Rayos X (muestra PDBXRay), la totalidad de las proteínas del PDB (PDBUniverso), y la base de datos de secuencias proteicas UniProtKB/Swiss-Prot, versión 6-Marzo-2013 (<http://web.expasy.org/docs/relnotes/relstat.html>), con 539.616 secuencias proteicas, y 191.569.459 aminoácidos (tabla 2.8.1.1.).

Tabla 2.8.1.1. Frecuencias de aminoácidos expresadas en porcentaje, en PDBGlobal, PDBXRay, PDBUniverso y Uniprot/SwissProt.

Aa	PDBGlobal	PDBXRay	PDBUniverso	SwissProt
C	1,26	1,27	1,39	1,37
W	1,37	1,31	1,40	1,08
M	2,34	2,36	2,35	2,42
H	2,81	2,78	2,67	2,27
Y	3,41	3,40	3,45	2,92
Q	3,62	3,66	3,74	3,93
F	3,89	3,93	3,94	3,86
N	4,23	4,24	4,28	4,00
P	4,64	4,58	4,70	4,70
R	4,91	4,95	4,92	5,53
I	5,54	5,69	5,62	5,96
T	5,57	5,54	5,64	5,34
K	5,66	5,72	5,74	5,84
D	5,86	5,78	5,70	5,45
S	6,06	6,17	6,23	6,56
E	6,52	6,67	6,59	6,75
V	7,06	7,03	7,02	6,87
G	7,87	7,61	7,58	7,08
A	8,53	8,30	8,09	8,25
L	8,82	9,01	8,93	9,66
Varianza	4,75	4,72	4,51	5,18
t-Student pareada		1,00	0,98	0,93

En la tabla 2.8.1.1., se muestran las frecuencias de los veinte aminoácidos estándar encontrados en proteínas, contados de forma individual, en las bases de datos del SwissProt y el Protein Data Bank (PDB, 2007, 2008). Se compararon las frecuencias de aminoácidos individuales entre el UniProtKB/Swiss-Prot, dos muestras del Protein Data Bank y la totalidad del PDB. Se presentan las varianzas de las tres muestras, y los valores de la prueba t-student para una confiabilidad del 95% ($\alpha=0.05$). Los valores muestran que la muestra PDBGlobal, es estadísticamente igual a las muestras PDBXRay, PDBUniverso y Uniprot/SwissProt, con lo que se puede considerar una muestra representativa del universo proteico representado en las bases de datos.

En la figura 2.8.1.1., se puede observar que las frecuencias de las tres muestras del PDB, son casi exactamente iguales, con las mayores diferencias en los aminoácidos Glicina (G), Alanina (A) y Leucina (L), que son los más frecuentemente hallados en proteínas. Sin embargo, estas diferencias no son estadísticamente significativas. Las mayores diferencias se presentan al comparar la muestra PDBGlobal con la totalidad de proteínas de la base de datos Uniprot/SwissProt, con más

de 500.000 proteínas, pero las diferencias no resultaron significativas. Con este resultado, decidimos entonces usar la distribución de nuestra propia muestra del PDB (PDBGlobal) en lugar de la del SwissProt.

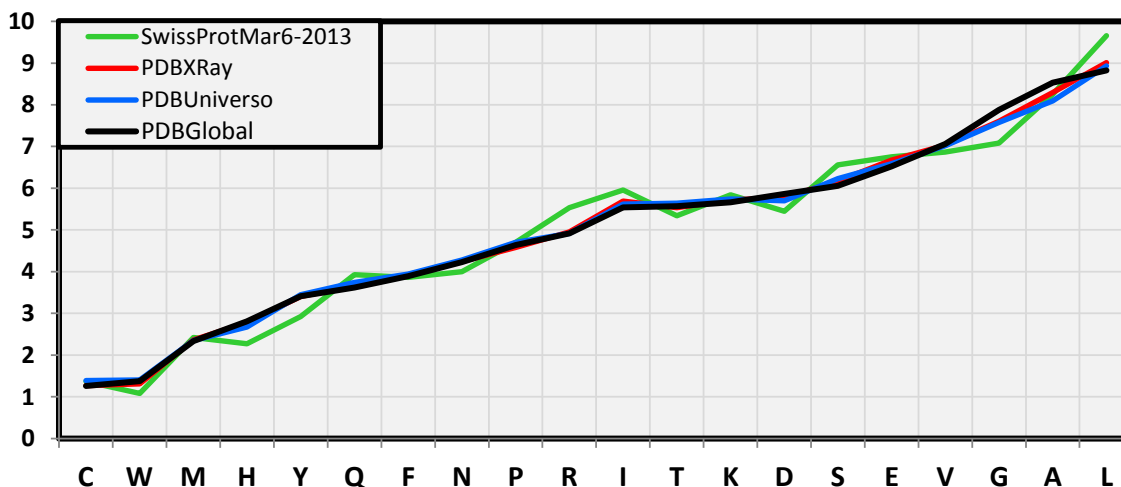


Figura 2.8.1.1. Comparación entre las frecuencias de aparición de aminoácidos en proteínas de PDBGlobal, PDBXRay, PDBUniversal y Uniprot/SwissProt.

En la figura 2.8.1.1., se observa que hay mucho parecido entre las frecuencias de los aminoácidos tomados de nuestra muestra del PDB (PDB Sample), y la totalidad de los aminoácidos encontrados en el SwissProt/Uniprot. Los aminoácidos Cisteína (C), Metionina (M), Fenilalanina (F), Asparragina (N), Prolina (P) Lisina (K) y Valina (V) presentan las menores diferencias en sus frecuencias entre las dos muestras (PDB vs SwissProt), con diferencias menores de 0.2; mientras tanto, los aminoácidos Histidina (H), Arginina (R), Serina (S), Glicina (G) y Leucina (L), presentan las mayores diferencias (mayores ó iguales a 0,5).

En las figura siguiente (figura 2.8.1.1.), el eje horizontal presenta los 20 aminoácidos convencionales en código de una letra, ordenados ascendentemente por su frecuencia; en el eje vertical, la frecuencia de aparición de cada aminoácido, expresada en porcentaje.

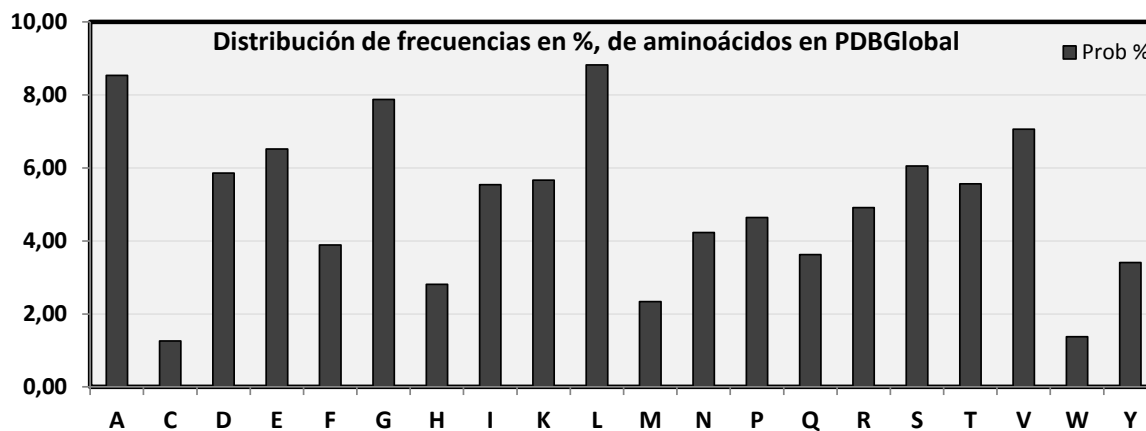


Figura 2.8.1.2. Distribución de las frecuencias de aminoácidos en proteínas de PDBGlobal.

Debido a que el número de proteínas de nuestra muestra PDBGlobal puede considerarse relativamente bajo con respecto a la totalidad del PDB, para verificar la representatividad de nuestra muestra, comparamos las frecuencias de los aminoácidos individuales de estas proteínas, con las de una muestra de todas las proteínas del PDB resueltas por Cristalografía de Rayos X (75215 proteínas. A este conjunto lo hemos llamado PDBXRay), y con la totalidad de las proteínas asentadas en el PDB hasta Junio 23 de 2013 (82893 proteínas; conjunto que llamamos PDBUniverso). Las frecuencias pueden verse en la tabla 2.8.2.1., lo mismo que las varianzas de las muestras, y las pruebas t que muestran que las muestras pueden considerarse iguales ó provenientes de la misma población, con una confianza del 95% ($\alpha=0.05$).

En la figura 2.8.1.1, se observa una coincidencia casi completa entre las tres curvas de frecuencias del PDB, y es evidente que las diferencias son muy pequeñas; se aplicó la prueba T, mostrando que los aminoácidos presentan las mismas frecuencias desde el punto de vista estadístico. Todo esto le da validez al uso de las frecuencias de aminoácidos de nuestra muestra, PDBGlobal.

La prueba t-student dice entonces que las muestra PDBGlobal, PDBXRay, PDBUniverso y Uniprot/SwissProt, provienen de la misma población y son estadísticamente iguales. Con este resultado, decidimos entonces usar la distribución de nuestra propia muestra PDBGlobal, en lugar de las PDBXRay y PDBUniverso.

2.8.2. Comparación entre PDBGlobal y PDBXRay en cuanto a las frecuencias de las estructuras secundarias

Para verificar que nuestra muestra PDBGlobal también es representativa en cuanto a su composición de estructuras secundarias, comparamos PDBGlobal y PDBXRay en dos aspectos:

1. La participación de cada aminoácido en cada estructura secundaria, y
2. La distribución de las estructuras secundarias en cada uno de los aminoácidos.

El primer análisis permite establecer si un aminoácido prefiere una estructura secundaria específica sobre las demás. El segundo análisis permite establecer si una estructura secundaria determinada muestra preferencia para algunos aminoácidos.

En el capítulo 3 se hace el análisis mas detallado de estas dos distribuciones en los diferentes grupos de proteínas. En el presente capítulo nos limitamos a mostrar, que no hay diferencia significativa en estos aspectos entre nuestra muestra PDBGlobal y los conjuntos más grandes PDBXRay y PDBUniverso.

2.8.2.1. La participación de cada aminoácido en cada estructura secundaria

La distribución de cada estructura secundaria sobre los 20 aminoácidos fue calculada como las veces que aparece un aminoácido en una estructura secundaria específica, dividido entre las veces que aparece esa estructura secundaria en la muestra proteica. Esta información nos dice en qué medida un aminoácido es más frecuente en una estructura secundaria, con respecto a los demás aminoácidos. Habla entonces de la composición de las estructuras secundarias en términos de los aminoácidos que las forman.

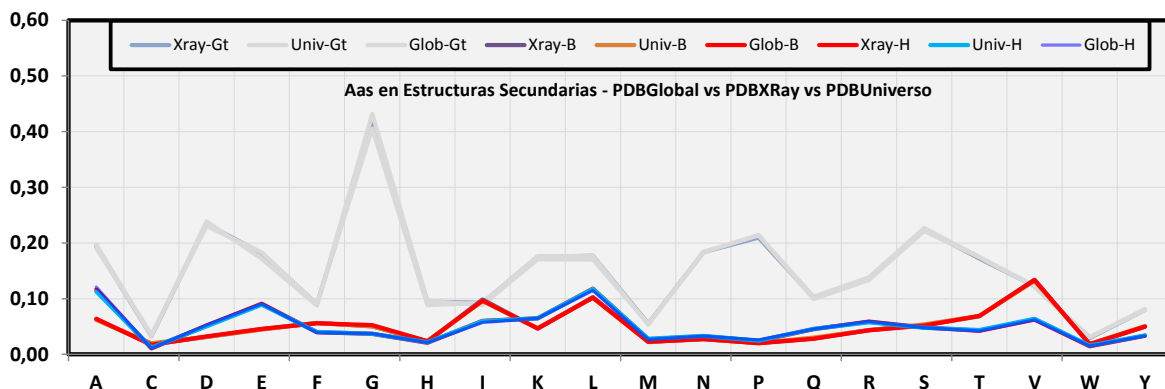


Figura.1.8.2.1.1. Participación de cada uno de los 20 aminoácidos en las estructuras secundarias clásicas: hélices, hojas y giros, en las muestras PDBGlobal, PDBXRay y PDBUniverso.

Como puede observarse en la figura 2.8.2.1.1., las frecuencias de aparición de los 20 residuos de aminoácido en las estructuras secundarias, son tan parecidas, que sólo se distinguen claramente, tres líneas, correspondientes a las estructuras helicoidales, hilos/hojas Beta y giros.

En la tabla 1.8.2.1.1., se muestran las frecuencias de aparición de los aminoácidos en las tres muestras, y debajo de todas, los valores de la prueba de t-student, por encima del umbral de 0.05, lo que confirma que las muestras provienen de la misma población y se pueden considerar estadísticamente iguales.

Tabla 2.8.2.1. Frecuencias de los aminoácidos en las estructuras secundarias DSSP-5 en PDBGlobal, PDBXRay y PDBUniverso, con los valores de la prueba t-student.

Aa	PDBGlobal					PDBXRay					PDBUniverso				
	B	H	S	T	U	B	H	S	T	U	B	H	S	T	U
A	0,17	0,49	0,06	0,09	0,19	0,17	0,49	0,06	0,09	0,20	0,16	0,48	0,06	0,09	0,21
C	0,31	0,29	0,07	0,08	0,24	0,31	0,29	0,07	0,07	0,25	0,31	0,29	0,07	0,08	0,25
D	0,13	0,31	0,12	0,15	0,29	0,12	0,31	0,12	0,14	0,31	0,12	0,30	0,12	0,15	0,31
E	0,16	0,48	0,07	0,11	0,17	0,15	0,47	0,07	0,11	0,20	0,15	0,46	0,08	0,11	0,20
F	0,33	0,35	0,06	0,07	0,18	0,31	0,35	0,07	0,07	0,19	0,31	0,35	0,07	0,07	0,20
G	0,15	0,16	0,16	0,27	0,26	0,14	0,17	0,16	0,25	0,27	0,14	0,17	0,17	0,25	0,27
H	0,19	0,26	0,08	0,10	0,37	0,18	0,26	0,08	0,10	0,39	0,19	0,28	0,08	0,10	0,35
I	0,40	0,36	0,04	0,04	0,16	0,38	0,37	0,05	0,04	0,17	0,37	0,36	0,05	0,04	0,18

K	0,19	0,40	0,08	0,12	0,21	0,18	0,39	0,08	0,12	0,23	0,18	0,38	0,09	0,12	0,23
L	0,26	0,45	0,05	0,06	0,17	0,25	0,45	0,05	0,06	0,18	0,25	0,45	0,05	0,07	0,19
M	0,22	0,39	0,05	0,06	0,27	0,21	0,40	0,05	0,06	0,28	0,21	0,41	0,05	0,06	0,27
N	0,15	0,27	0,11	0,19	0,28	0,14	0,27	0,12	0,18	0,29	0,14	0,26	0,12	0,18	0,29
P	0,10	0,19	0,09	0,17	0,44	0,10	0,18	0,10	0,16	0,46	0,09	0,18	0,10	0,17	0,46
Q	0,18	0,44	0,07	0,10	0,21	0,17	0,43	0,07	0,10	0,23	0,17	0,42	0,08	0,10	0,23
R	0,20	0,42	0,08	0,10	0,20	0,19	0,41	0,08	0,10	0,22	0,19	0,40	0,08	0,10	0,23
S	0,20	0,27	0,10	0,11	0,31	0,19	0,27	0,10	0,11	0,33	0,19	0,26	0,11	0,12	0,33
T	0,29	0,26	0,10	0,09	0,27	0,27	0,27	0,09	0,08	0,28	0,26	0,26	0,10	0,09	0,29
V	0,43	0,30	0,05	0,04	0,17	0,42	0,31	0,05	0,04	0,19	0,41	0,31	0,05	0,04	0,19
W	0,31	0,38	0,06	0,08	0,17	0,30	0,39	0,06	0,08	0,18	0,29	0,39	0,07	0,08	0,18
Y	0,34	0,34	0,07	0,08	0,18	0,32	0,34	0,07	0,08	0,19	0,32	0,34	0,07	0,08	0,19
Valores de la prueba t pareada						1,00					1,00				

Como puede observarse en las figuras 1.8.2.1.1. y 1.8.2.1.2., las frecuencias de aparición de aminoácidos en las diferentes estructuras secundarias consideradas en el presente trabajo (Hélices, Estructuras Beta, y Giros; estos últimos a su vez descompuestos en Giros tipo dobléz, Giros Beta y Giros al azar), son estadísticamente (y visualmente según las figuras) al compararse las frecuencias tomadas de la muestra PDBGlobal, de la muestra PDBXRay y de la totalidad de proteínas del PDB, en PDBUniverso.

En la figura 2.8.2.1.2. se han dividido los giros en dobleces, giros Beta y random coils, lo que permite ver la participación de cada aminoácido en cada estructura secundaria.

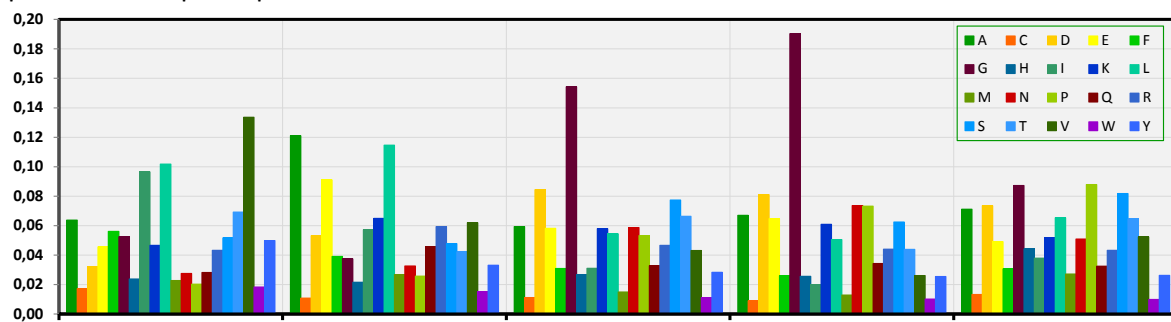


Figura 2.8.2.1.2. Participación de cada uno de los 20 aminoácidos en las estructuras secundarias DSSP-5, en la muestra PDBGlobal.

Como puede observarse en las figuras 1.8.2.1.1. y 1.8.2.1.2., las frecuencias de aparición de aminoácidos en las diferentes estructuras secundarias (Hélices, Estructuras Beta, y Giros; estos últimos a su vez descompuestos en Giros tipo dobléz, Giros Beta y Giros al azar), son estadísticamente iguales (y visualmente según las figuras) al compararse las frecuencias tomadas de la muestra PDBGlobal, de la muestra PDBXRay y de la totalidad de proteínas del PDB, en PDBUniverso.

2.8.2.2. La distribución de las estructuras secundarias en cada uno de los aminoácidos

La distribución de las estructuras secundarias en cada uno de los aminoácidos es calculada como el número de veces que aparece un aminoácido en una estructura secundaria específica, dividido entre el número de veces que aparece ese aminoácido en la totalidad de la muestra. Esta información nos dice en qué medida, cada aminoácido asume una estructura secundaria específica. Habla entonces de la tendencia de cada aminoácido a formar estructuras secundarias específicas. En la figura 2.8.2.2.1., se observa esta distribución, que se discutirá más exhaustivamente en el capítulo 3.

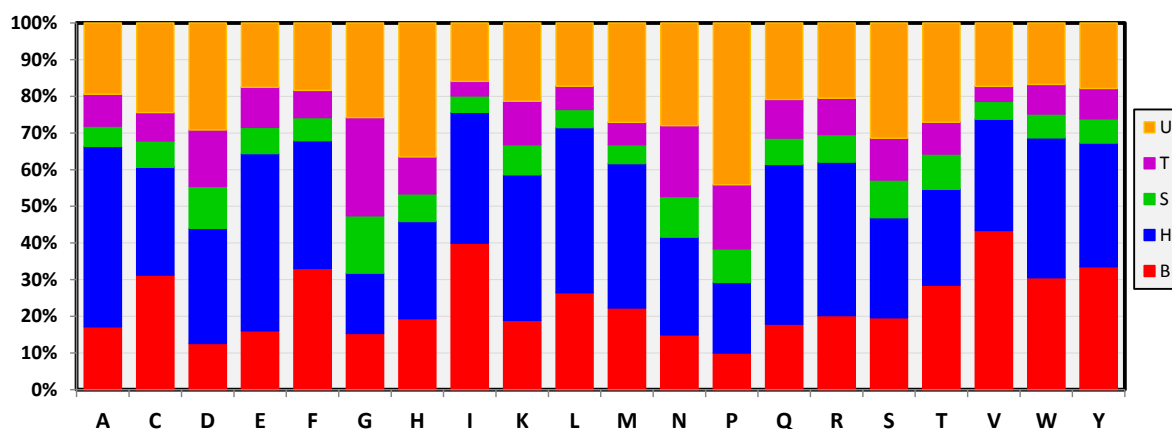


Figura 2.8.2.2.1. La distribución de las estructuras secundarias en cada uno de los aminoácidos en la muestra PDBGlobal.

CONCLUSIONES CAPÍTULO 2 - SELECCIÓN DE LA MUESTRA

1. El conjunto PDBGlobal es representativo de la totalidad de PDB y del universo representado en él, en cuanto a la frecuencia de aparición de aminoácidos, la distribución de cada estructura secundaria sobre los 20 aminoácidos, la distribución de las estructuras secundarias en cada uno de los aminoácidos, y la distribución de tripletas de residuos, junto con sus estructuras secundarias.
2. Por lo anterior, se justifica la utilización de la muestra PDBGlobal, como nuestra muestra de referencia para las comparaciones de frecuencias de aminoácidos individuales, frecuencias de aminoácidos en estructuras secundarias, propensiones de Chou y Fasman (capítulo 2) y el análisis sobre tripletas de aminoácidos y sus correspondientes tripletas estructurales que generalizan estos conceptos (capítulo 3).
3. Esta muestra, y las submuestras resultantes de la agrupación por criterios como contenido de estructura secundaria, localización celular, organismo del que provienen, Phylum al que pertenecen los organismos, Tipo de Enzima según la Enzyme Commission y Familia proteica, se usarán en trabajos posteriores a este proyecto de investigación.

BIBLIOGRAFÍA CAPÍTULO 2 - SELECCIÓN DE LA MUESTRA

Andreeva A., Howorth D., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2004). **SCOP database in 2004: refinements integrate structure and sequence family data**. Nucl. Acid Res. 32:D226-D229.

Andreeva A., Howorth D., Chandonia J.-M., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2007). **Data growth and its impact on the SCOP database: new developments**. Nucl. Acid Res. advance access, doi:10.1093/nar/gkm993.

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS (January 2004). **"UniProt: the Universal Protein knowledgebase"**. Nucleic Acids Res. 32 (Database issue): D115–9.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L., Studholme, D.J., Yeats C. y Eddy S.R. 2004. **The Pfam Protein Families Database**. Nucleic Acids Research, 32:D138-D141.

Bateman, A. Birney, E., Cerruti, L. Durbin, R. Ewiler, L. Eddy, S.R. Griffiths-Jones, S. Howe, K.L. Marshall, M. y Sonnhammer, E.L. **The Pfam Protein Families Database**. Nucleic Acids Research (2002) 30(1):276-280.

Bateman, A. Birney, E., Durbin, R., Eddy, S.R., Howe K.L. y Sonnhammer E.L. **The Pfam Protein Families Database**: Nucleic Acids Research (2000) 28:263-266.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R. Finn R.D. y Sonnhammer, E.L.L. **Pfam 3.1: 1313 multiple alignments match the majority of proteins**. Nucleic Acids Research (1999) 27:260-262.

Finn R.D. **Pfam: the protein families database**. (eds M.J. Dunn, L.B. Jorde, P.F.R. Little, S. Subramaniam) en: **Genetics, Genomics, Proteomics and Bioinformatics, Section 6: Protein Families** (2005) ISBN 978-0-470-84974-3 2005.

Finn, J. Mistry, B. Schuster-Böckler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S.R. Eddy, E.L.L. Sonnhammer and A. Bateman **Pfam: clans, web tools and services**. Nucleic Acids Research (2006) Database Issue 34:D247-D51.

Finn, R.D. Bateman, A. y Griffiths-Jones, S. **Identifying protein domains with the Pfam database**. Current protocols in bioinformatics ISBN 978-0-471-25093-7.

Lo Conte L., Brenner S. E., Hubbard T.J.P., Chothia C., Murzin A. (2002). **SCOP database in 2002: refinements accommodate structural genomics**. Nucl. Acid Res. 30(1), 264-267.

Moss, G.P. 1992. **Enzyme Nomenclature 1992**. Academic Press, San Diego, California, ISBN 0-12-227164-5, 0-12-227165-3. Suplemento1 (1993), Suplemento2 (1994), Suplemento3(1995), Suplemento4 (1997) y Suplemento5 (en Eur.J.Biochem. 1994, 223, 1-5; Eur.J.Biochem. 1995, 232, 1-6; Eur.J.Biochem. 1996, 237, 1-5; Eur.J.Biochem. 1997, 250, 1-6, y Eur.J.Biochem.1999, 264, 610-650; respectivamente) [Copyright IUBMB].

Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). **SCOP: a structural classification of proteins database for the investigation of sequences and structures**. *J. Mol. Biol.* 247, 536-540.

Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of **proteins database for the investigation of sequences and structures**. *J. Mol. Biol.* 247, 536-540.

Oshima, T., Imahori, K. (1974). **Description of *Thermus thermophilus* (Yoshida and Oshima) comb. nov., a nonsporulating thermophilic bacterium from Japanese thermal spa**. *Int J Syst Bacteriol.* 24:102–112

Webb EC., (1992). **Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes**. San Diego: Published for the International Union of Biochemistry and Molecular Biology by Academic Press.

CAPÍTULO 3. FRECUENCIAS DE AMINOÁCIDOS Y TRIPLETS

En el presente capítulo se obtuvieron y analizaron las frecuencias de aparición de aminoácidos individuales en la muestra representativa del PDB (PDBGlobal, definido en el capítulo anterior) y en proteínas agrupadas según diferentes criterios, como contenido de estructura secundaria según la clasificación SCOP, localización celular, organismo en el que se encontraron, Phylum al que pertenecen los organismos en que se encontraron las proteínas, tipo de enzima según la Comisión de Enzimas (EC), y familia proteica según PFam. En todos los casos, se analizaron las frecuencias de los veinte aminoácidos estándar hallados en proteínas y codificados en el código genético (con la excepción de Pirrolisina y Selenocisteína, por estar muy poco representados en el PDB), en cada grupo de proteínas seleccionado. Todos los resultados se contrastaron con los resultados correspondientes obtenidos de todo el conjunto PDBGlobal.

En cada caso, tanto en aminoácidos individuales, como en triplets, se contrastan y analizan las frecuencias, con las frecuencias de un alfabeto reducido de propiedades fisicoquímicas de los aminoácidos, tal como se describe en la metodología y en el capítulo 2.

Con base en las frecuencias de aparición en proteínas, se identificaron los grupos de proteínas en que aparecen aminoácidos con frecuencias claramente diferentes a la muestra representativa de las proteínas encontradas en el PDB, y se discutió sobre la posible explicación biológica para que esto ocurra. También se analiza su potencial como herramienta para el rediseño de proteínas y la ingeniería de las mismas.

3.1. Frecuencias de residuos de aminoácidos en las proteínas agrupadas por los criterios seleccionados.

Se analizaron las frecuencias de los veinte aminoácidos tanto en la muestra de la generalidad de las proteínas, como de las proteínas agrupadas por los criterios señalados antes. Además de ello, se redujeron los veinte aminoácidos a un alfabeto reducido de nueve letras, que agrupa a los aminoácidos según algunas propiedades fisicoquímicas. Con base en este alfabeto reducido, también se calcularon las frecuencias de las nueve letras, que representan propiedades fisicoquímicas. Con esto, se hace un doble análisis, tanto en términos de los aminoácidos individuales, como en términos de algunas propiedades fisicoquímicas.

Los veinte aminoácidos quedaron entonces agrupados en las siguientes nueve letras del alfabeto reducido, representando propiedades, de la siguiente manera:

Se dejaron sin agrupar los aminoácidos que presentan propiedades particulares en las proteínas:

Glicina (G), por ser el aminoácido más pequeño, y permitir el máximo empaquetamiento, gracias a su cadena lateral. En las propensiones de Chou y Fasman, se encuentra formando giros en proteínas; Prolina (P), por ser el único aminoácido que puede formar enlaces tanto trans como cis, y por la distorsión que causa su grupo amino secundario en la formación del enlace peptídico. Triptófano (W), por ser el residuo de aminoácido de mayor tamaño, ser el de menor frecuencia en muchas proteínas, producido por un solo codón, y su carácter apolar. Cisteína, por ser el único aminoácido formador de puentes de disulfuro, y por tanto, ser determinante en la estabilización estructural de muchas proteínas. Así, el alfabeto reducido de los veinte aminoácidos, quedó de la siguiente manera:

Tabla 3.1.1. Alfabeto reducido para los veinte aminoácidos (aas), en un alfabeto de 9 letras, correspondiente a propiedades.

Criterio para agrupar aminoácidos	Aminoácidos	Letra en el nuevo Alfabeto reducido
Grupo de residuos de aminoácidos hidrofóbicos.	A, F, I, L, M, V	U
Grupo de aminoácidos con cadena lateral básica (con carga +), muy polares	H, K, R	B
Grupo de aminoácidos con cadena lateral –OH (forma enlaces de H)	S, T, Y	O
Grupo de aminoácidos con cadena lateral ácida (con carga negativa), muy polares	D, E	X
Grupo de aminoácidos con cadena lateral Amida	N, Q	Z
Aminoácido formador de puentes de disulfuro	C	C
Aminoácido con cadena lateral más pequeña (H), que permite acercamiento entre cadenas, y la mayor libertad de giros en los ángulos Phi y Psi.	G	G
Iminoácido, ocasiona distorsiones en estructuras secundarias regulares por formar un enlace peptídico de un amino secundario	P	P
Aminoácido con la cadena lateral más grande. Costoso transportarlo en citoplasma, el menos frecuente en proteínas, codificado por un solo codón.	W	W

3.3.1.0. Frecuencias de aminoácidos en proteínas de la muestra PDBGlobal

En esta sección, se estudiaron las frecuencias de residuos de aminoácidos en la muestra PDBGlobal, en 19.308 proteínas, que sumaron 8'862.961 aminoácidos. A continuación, se presentan las frecuencias de los aminoácidos en PDBGlobal, y se puntualizan las tendencias generales y las diferencias más importantes.

Tabla 3.1.0.1. Frecuencias de los veinte aminoácidos estándar en proteínas de la muestra PDBGlobal

Aa	Frec	Prob %
A	756356	8,53
C	111760	1,26
D	519677	5,86
E	578065	6,52
F	344695	3,89
G	697945	7,87
H	248992	2,81
I	491196	5,54
K	502015	5,66
L	782124	8,82
M	207158	2,34
N	375054	4,23
P	411155	4,64
Q	321040	3,62
R	435595	4,91
S	536874	6,06
T	493484	5,57
V	625951	7,06
W	121758	1,37
Y	302067	3,41
	8862961	100

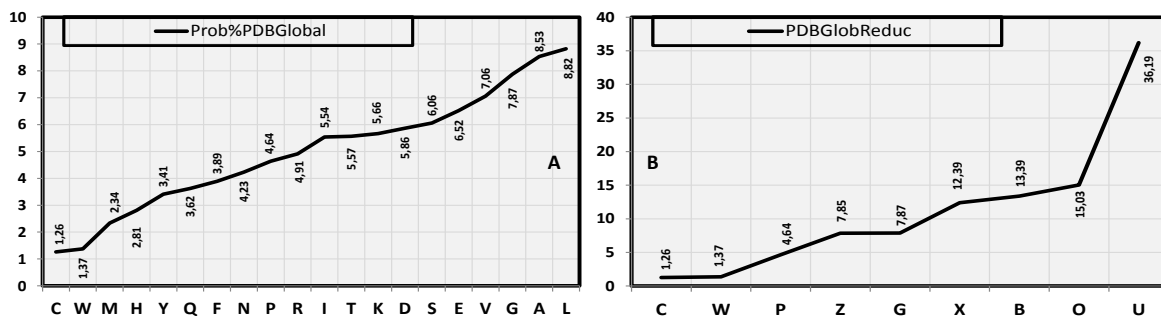


Figura 3.1.0.1. Frecuencias de aminoácidos individuales en PDBGlobal, organizadas en orden ascendente por frecuencia.

Tendencias generales y Diferencias importantes

Como puede verificarse en la tabla 3.1.0.1. y en figura 3.1.0.1., en PDBGlobal, el residuo de aminoácido menos frecuentemente encontrado en proteínas, es la Cisteína (C), con una frecuencia en la muestra PDBGlobal, de 1.26%. Este valor se podría considerar como la frecuencia de la Cisteína en la generalidad de las proteínas, aunque a lo largo de este trabajo, se verá que este aminoácido presenta variaciones importantes en algunos grupos de proteínas, con lo que tomará importancia como indicador de la estabilización de una proteína mediante enlaces fuertes, o enlaces débiles.

La mayor importancia de la presencia de las Cisteínas en una proteína, corresponde al potencial de formar puentes de disulfuro, estabilizando la estructura terciaria y/o cuaternaria de la proteína

(Gilbert, 1990; Gilbert, 1995; Tu y Weissman, 2004; Witt, 2008). De allí que el contenido de Cisteínas en una proteína o un grupo de ellas, puede hablar de si la estabilización de la estructura 3D se debe a fuerzas débiles o a fuerzas fuertes, lo que a su vez, puede ofrecer indicios sobre si la proteína puede perder fácilmente o no, su plegamiento nativo.

En el PDBGlobal, la mayoría de las proteínas se ha aislado de ambientes parcialmente reductores. A manera de ejemplo, la mitad de las proteínas del PDB, corresponde a proteínas de Citoplasmas eucariotes. En general, el citoplasma eucariote es un ambiente parcialmente reductor, por lo que es difícil la formación de enlaces de disulfuro y por tanto las proteínas que permanecen en ese ambiente, tienden a tener bajos contenidos de Cisteína. La función principal de este aminoácido es la formación de puentes disulfuro que estabilizan la estructura tridimensional y/o la unión entre subunidades (Gilbert, 1990; Gilbert, 1995; Tu y Weissman, 2004; Witt, 2008).

El segundo residuo de aminoácido menos frecuente en las proteínas del PDBGlobal, es el Triptófano (W), con una frecuencia de 1.37% (Tabla 3.1.0.1. y figura 3.1.0.1.). Este aminoácido está codificado por un solo codón (UGG), es el que presenta la cadena lateral más voluminosa, que puede significar un mayor costo energético al ser transportado en el citoplasma. Ya desde su descubrimiento por Frederick Hopkins in 1901, al hidrolizar 600 g de Caseína, y obtener sólo 4g de Triptófano (Hopkins, 1901), evidenciaba la baja cantidad de éste residuo en proteínas.

Sólo la Cisteína y el Triptófano presentan frecuencias menores a 2%. La Metionina y la Histidina se encuentran en frecuencias de entre 2 y 3% en las proteínas de PDBGlobal. Tirosina, Glutamina y Fenilalanina, presentan frecuencias de entre 3 y 4%. Asparragina, Prolina, Arginina, Isoleucina, Trenonina, Lisina, y Ácido Aspártico, aparecen en las frecuencias medias, entre 4 y 6%. Serina, y Ácido Glutámico están entre 6 y 7% en las proteínas, mientras que los residuos más frecuentes, con comportamientos hidrofóbicos, son Valina, Glicina, Alanina y Leucina, entre 7 y 9%.

El alto contenido de residuos hidrofobicos en proteínas, podría explicarse en que estos cuatro aminoácidos tienden a aparecer en el corazón hidrofóbico de las proteínas, aún de las proteínas que están inmersas en un medio acuoso. La hidrofobicidad es un criterio suficiente para construir un corazón proteico para las enzimas (Axe et al., 1996).

En la figura 3.1.0.1.B, se presentan las frecuencias del alfabeto reducido de propiedades de aminoácidos: C, W, P, Z, G, X, B, O, y U. En esta figura se observa que los aminoácidos hidrofóbicos, agrupados en la letra “U”, resultan ser la agrupación más frecuente, apareciendo en el 36.19% de las proteínas. Estos aminoácidos estarían formando parte de la superficie de proteínas de membrana, y en general, de todas aquellas que interactúen en ambientes no polares; y también del corazón hidrofóbico de proteínas globulares que permanecen en ambientes acuosos, como el citoplasma. Este corazón hidrofóbico permanece usualmente inaccesible al solvente (Tsai et al, 1997), y puede explicar parcialmente el plegamiento de proteínas, gracias a la fuerza generada por las interacciones hidrofóbicas. La frecuencia de los residuos hidrofóbicos, suma más del doble de la frecuencia siguiente más alta, la de los residuos que tienen un hidroxilo en su cadena lateral (S,

T, H), agrupados en la letra “O”; estos residuos con un grupo hidroxilo, son importantes en la superficie de las proteínas, pues son blanco de algunos tipos de modificaciones postraduccionales, que constituyen un alto porcentaje de las proteínas (Mann y Jensen, 2003; Seo y Lee, 2004).

Los aminoácidos básicos, agrupados en la letra “B”, presentan una frecuencia total de aparición en PDBGlobal, de 13.39%. Los aminoácidos ácidos D, y E, agrupados en la letra “X”, suman una frecuencia de 12.39%. Los aminoácidos con cadena lateral amida, agrupados en la letra “Z”, suman una frecuencia menor que la del aminoácido Glicina. Las frecuencias de los demás aminoácidos individuales que no se agruparon, son obviamente las mismas que en la parte A de la gráfica.

3.1.1. Frecuencias de aminoácidos de las proteínas clasificadas por su composición de estructura secundaria, según SCOP.

En el presente capítulo se obtuvieron y analizaron las frecuencias de aparición de los aminoácidos individuales en la proteínas agrupadas por su contenido de estructura secundaria según la Clasificación Estructural de Proteínas (SCOP por su sigla en inglés) (Murzin, et al, 1995; Lo Conte et al, 2002; Andreeva et 2004; Andreeva et al, 2007). Se compararon las frecuencias de los veinte aminoácidos estándar hallados en proteínas y codificados en el código genético, de cada grupo de proteínas seleccionado, con la muestra global PDBGlobal. A continuación, se presentan las frecuencias de aminoácidos individuales en las proteínas clasificadas según su composición de estructura secundaria, de acuerdo con SCOP.

Tabla 3.1.1.1. Frecuencias de aparición de los veinte aminoácidos estándar en las proteínas agrupadas según su composición de estructura secundaria según la clasificación SCOP.

Aa	PDBGlobal	AllAlpha	AllBeta	Alpha-Beta	Alpha+Beta	CoiledCoil	Small	Multidomain	Others
C	1.26	1.27	1.36	1.16	1.21	1.16	5.64	1.14	0.78
W	1.37	1.21	1.72	1.22	1.31	1.22	1.49	1.22	2.69
M	2.34	2.56	2.08	2.14	2.14	2.14	1.84	2.39	1.76
H	2.81	2.59	2.62	2.56	2.73	2.56	2.73	2.27	1.73
Y	3.41	3.34	3.79	3.28	3.44	3.28	3.77	3.51	3.88
Q	3.62	3.98	3.64	3.38	3.55	3.38	3.58	3.51	4.25
F	3.89	3.86	4.02	3.87	3.91	3.87	3.81	3.86	3.73
N	4.23	3.94	4.71	3.96	4.07	3.96	4.47	3.91	5.66
P	4.64	4.17	4.98	4.64	4.52	4.64	5.72	4.91	4.47
R	4.91	5.17	4.55	5.03	5.17	5.03	5.45	5.25	3.79
I	5.54	5.50	5.31	5.94	5.80	5.94	4.60	5.95	5.32
T	5.57	5.03	6.26	5.18	5.30	5.18	5.40	5.37	5.61
K	5.66	6.32	5.57	5.74	5.89	5.74	5.99	6.25	4.99
D	5.86	5.84	5.92	5.77	5.82	5.77	5.23	5.78	4.92
S	6.06	5.71	6.82	5.54	5.75	5.54	6.67	5.15	6.92
E	6.52	7.56	6.01	6.80	7.19	6.80	6.75	7.29	5.61
V	7.06	6.40	7.44	7.45	7.20	7.45	5.82	7.12	7.26
G	7.87	6.66	8.27	7.94	7.58	7.94	7.58	7.61	8.33
A	8.53	9.06	7.19	9.06	8.28	9.06	6.17	8.58	9.32
L	8.82	9.83	7.75	9.01	8.84	9.01	7.30	8.92	9.02

Tendencias generales y Diferencias importantes

Se realizó una prueba de t-student para muestras pareadas, entre el PDBGlobal y cada uno de los grupos de proteínas, y se encontró que todos los grupos presentan diferencias significativas en sus frecuencias, con respecto a las del PDBGlobal. Esto indica que estas agrupaciones de proteínas presentan composiciones diferentes de residuos de aminoácidos. En esta sección analizamos las particularidades y generalidades de las diferencias y las similitudes.

Si se considera la máxima diferencia entre la frecuencia de cada aminoácido en cada grupo de proteínas y la frecuencia en PDBGlobal (tabla 3.1.1.2), se puede observar que el grupo de proteínas que presentó la máxima diferencia en un mayor número de aminoácidos, es el grupo “Others”, que corresponde a proteínas que por su contenido de estructura secundaria no se pudieron clasificar dentro de los demás grupos proteicos. Algunas de estas proteínas corresponden a proteínas intrínsecamente desordenadas, al igual que algunas proteínas clasificadas como “small”.

Sobre esas estructuras y su contenido de estructura secundaria, se discutirá en el capítulo 3. Estas diferencias sugieren que los aminoácidos con los que se construyen las estructuras al azar, son diferentes que los que se usan para construir estructuras secundarias regulares, del tipo hélices u hojas. En este grupo de proteínas “Others”, las diferencias máximas con respecto a PDBGlobal ocurren en los siguientes aminoácidos: Fenilalanina (F) (0.16% de diferencia con la frecuencia de Fenilalanina (F) en PDBGlobal), Tirosina (Y) (0.47%), Metionina (M) (0.58%), Glutamina (Glutamina (Q)) (0.63%), Lisina (K) (0.68%), Histidina (H) (1.08%) Arginina (R) (1.12%), Triptófano (W) (1.32%) y Asparagina (N) (1.42%). La suma de todas estas diferencias, es de 8.41%. Es el grupo de proteínas clasificadas por su contenido de estructura secundaria, que más aminoácidos presenta con diferencias máximas en sus frecuencias con respecto a PDBGlobal.

Sin embargo, el grupo de proteínas que presenta una mayor sumatoria de las diferencias, es el grupo de proteínas clasificadas como “Small” en SCOP, que las define como “dominadas por ligandos metálicos, grupos hemo, ó puentes de disulfuro” (<http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.b.html>. consultado el 30 de Junio de 2013). Las diferencias más grandes entre las frecuencias de proteínas “Small”, con las del PDBGlobal, son: Serina (S) (0.90%), Isoleucina (I) (0.95), Pro (1.08), Valina (V) (1.24%), Leucina (L) (1.53%), Alanina (A) (2.36%) y Cisteína (C) (4.38), con un acumulado de diferencias del 15.75%. Esto puede observarse en la tabla 3.1.1.2. y su figura correspondiente (2.1.1.1.).

En la tabla 3.1.1.2., se observa que también en las proteínas TodoAlfa, el Ácido Glutámico (E), presenta diferencias importantes con el PDBGlobal. Esto puede ser explicado con el hecho de que aparentemente este residuo de aminoácido es un fuerte formador de estructuras helicoidales, y en estas proteínas con alto contenido de hélices, estará sobrerrepresentado.

Tabla 3.1.1.2. Diferencias máximas entre las frecuencias del PDBGlobal y las proteínas clasificadas por su contenido de estructura secundaria según SCOP.

Aa	MaxDiferencia	Tipo Proteína
F	0.16	Others
Y	0.47	Others
M	0.58	Others
Q	0.63	Others
K	0.68	Others
T	0.69	AllBeta
S	0.90	Small
I	0.95	Small
D	0.95	Others
E	1.04	AllAlpha
P	1.08	Small
H	1.08	Others
R	1.12	Others
G	1.21	AllAlpha
V	1.24	Small
W	1.32	Others
N	1.42	Others
L	1.53	Small
A	2.36	Small
C	4.38	Small

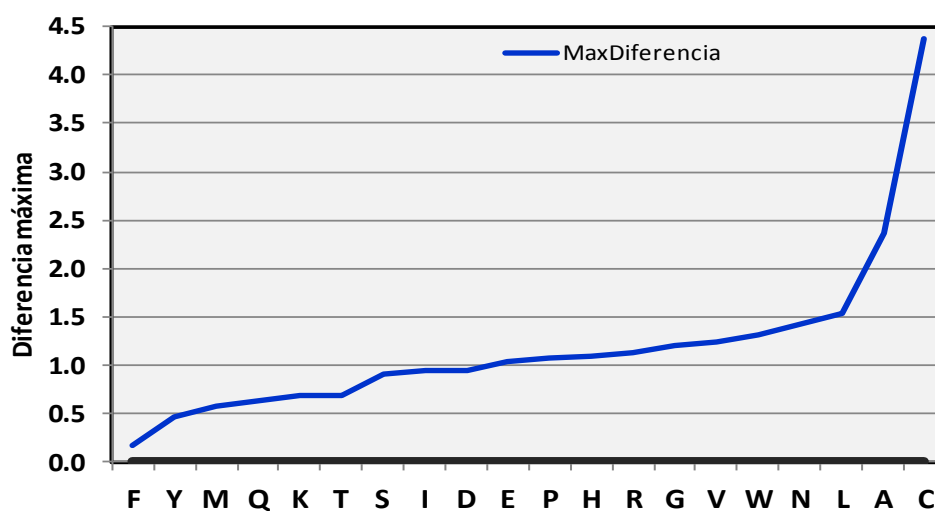


Figura 3.1.1.1. Magnitudes de las máximas diferencias entre las frecuencias del PDBGlobal, y las frecuencias de las proteínas agrupadas según composición de estructura secundaria, según SCOP.

La línea negra en la diferencia cero, representa al PDBGlobal.

De acuerdo con la figura 3.1.1.1., el aminoácido Triptófano, que constituye el 1.37% de las proteínas en PDBGlobal, en las proteínas clasificadas como “Otras”, aparece casi el doble de veces. Este resultado es llamativo, dado que la frecuencia de aparición del Triptófano en proteínas, es

casi la mínima de un aminoácido en proteínas. Las proteínas clasificadas como “otras”, no presentan los contenidos de estructura secundaria de las demás proteínas de SCOP; es decir, tendrán componentes de estructura secundaria diferentes, y por tanto composición de aminoácidos diferentes. No se encontró referencia a esto en la literatura, y es probable que sea una observación nueva para este residuo de aminoácido.

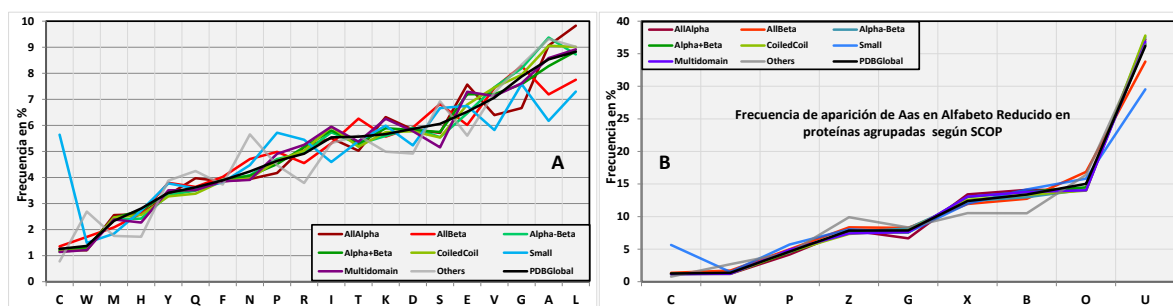


Figura 3.1.1.2. Frecuencias de aparición de los veinte aminoácidos estándar en las proteínas agrupadas según su composición de estructura secundaria según la clasificación SCOP.

Entre otras de las observaciones más conspicuas está el hecho de que la Cisteína (C), que usualmente es uno de los aminoácidos menos frecuentemente encontrado en proteínas (presente sólo en el 1.26% de las proteínas de PDBGlobal), aparecen con casi 4.5 veces más en las proteínas “pequeñas”: Es decir, estas proteínas pequeñas usan mucho más el aminoácido Cisteína para construirse, que la globalidad de las proteínas, lo que posiblemente está implicando que estas proteínas pequeñas requieren de más puentes de disulfuro para estabilizar su estructura tridimensional, que el resto de las proteínas. Probablemente, al ser más pequeñas, cuentan con un menor número posible de enlaces débiles (como enlaces de H), para estabilizar su estructura, con lo que los enlaces de disulfuro ganan importancia para mantener la estabilidad de la estructura tridimensional. Estas proteínas efectivamente se encuentran mucho menos estabilizadas por enlaces de hidrógeno ó por interacciones hidrofóbicas (Cheek et al., 2006), y entre ellas se encuentran Factores de crecimiento, toxinas, inhibidores enzimáticos, hormonas, feromonas, y alergenicos.

El aminoácido Prolina (P), es también más abundante en estas proteínas pequeñas, lo que podría estar explicado en el hecho de que estas proteínas presentan una cantidad relativamente alta de giros con respecto al total de contenido de estructura secundaria, y sus hélices e hilos Beta, tienden a ser más cortos. Por esto, al tener hélices e hilos y hojas más cortos, la cantidad relativa de los giros, se ve aumentada.

También es llamativo que en las proteínas pequeñas, los aminoácidos Leucina (L) y Alanina (A) tengan menores frecuencias (Leucina (L), poco más del 7% y Alanina (A) poco más del 6%) (Capítulo 1) las que presentan en la mayoría de las proteínas. En general, estos son los aminoácidos más frecuentes en todas las proteínas, con frecuencias de alrededor del 10%. Estos dos aminoácidos se consideran fuertes formadores de hélice Alfa, lo que encuentra una

explicación en el hecho de que en las proteínas pequeñas, las estructuras secundarias regulares, como hélices y hojas, son muy cortas ó inexistentes.

Mientras tanto, la Glicina es el aminoácido más frecuentemente encontrado en ellas (en proteínas “pequeñas”), lo que puede ser explicado por el hecho de que estas proteínas presenten un contenido de giros relativamente más alto que otras proteínas (Capítulo 1), y un alto nivel de empaquetamiento, dado que la Glicina es el residuo de aminoácido cuya cadena lateral permite la mayor cercanía con otras zonas de la cadena.

El aminoácido más frecuente en las proteínas Todo-Alfa, Alfa-Beta, proteínas Multidominio y Alfa+Beta, es la Leucina, y el segundo más frecuente en ellas es la Alanina (A) lo cual es consistente con la globalidad de las proteínas y el PDBGlobal. Sin embargo, en las Todo-Beta, Pequeñas proteínas, y Otras (que corresponden a proteínas que no presentan una estructura secundaria regular clasificada en las otras categorías SCOP), el residuo más frecuente es la Glicina, posiblemente debido a la necesidad de aumentar el número de giros o de vueltas sin estructura secundaria anexa, o de empaquetar la estructura proteica. La Ala y Leucina (L) son aminoácidos considerados propensos a formar hélices, lo cual es consistente con sus frecuencias más bajas en proteínas Todo-Beta.

Las proteínas Todo-Alfa presentan frecuencias más altas que las demás proteínas, en los residuos de Glutámico (E), Ala y Leucina (L), típicos formadores de hélices, mientras que frecuencias más bajas en los residuos Valina (V) y Glicina (G), más frecuentemente encontrados en hilos Beta y en giros, respectivamente.

Las proteínas Todo-Beta presentan frecuencias más bajas en los aminoácidos conocidos como propensos a formar hélices, como Glutámico (E), Ala y Leucina (L).

Las proteínas clasificadas como “Otras”, son ricas en estructuras al azar (“random coils”), son las proteínas que presentan la más alta frecuencia en Glicina (G), y que además es su aminoácido más frecuente. Glicina (G) es un aminoácido fuerte formador de giros y además, el aminoácido que más libertad de giro (en los ángulos Phi y Psi, de Ramachandran) y por tanto es de esperar que aparezca con frecuencia en estructuras al azar.

En la figura 3.1.1.2.B., es evidente que las proteínas agrupadas por su contenido de estructura secundaria, siguen aproximadamente las frecuencias de la muestra PDBGlobal, con pequeñas diferencias en las proteínas clasificadas como “small”, que presentan menor contenido de residuos apolares, y en las clasificadas como “others”, con menor contenido de aminoácidos ácidos.

Además de lo anterior, vale considerar que las diferencias en frecuencia de aminoácidos en un grupo proteico específico, con respecto a las frecuencias encontradas en el PDB, pueden estar asociadas con la magnitud de la frecuencia. Es decir, en aminoácidos frecuentes, diferencias grandes, pueden ser en realidad pequeñas diferencias relativas, y en residuos con frecuencias

pequeñas, las diferencias pueden no evidenciarse en su real magnitud. Así, las pequeñas diferencias encontradas en las frecuencias de Cisteína, se perciben como poco notorias en comparación con las diferencias de los aminoácidos con valores de frecuencia más altos, Leucina. Debido a esto, presentamos una figura con una normalización de las diferencias, que hemos llamado “Diferencias relativas” de manera que cada una de ellas, se pueda observar en proporción a la frecuencia del aminoácido encontrado en PDBGlobal. Esto magnificó las diferencias relativas de los aminoácidos poco frecuentes, y minimizo un poco las diferencias de los aminoácidos muy frecuentes. Presentamos una figura de las diferencias relativas, en la figura 3.1.1.3.

Tabla 3.1.1.3. Diferencias normalizadas entre las frecuencias del grupo de referencia (PDBGlobal) y las proteínas clasificadas según SCOP.

Aa	PDBGlobal	AllAlpha	AllBeta	Alpha-Beta	Alpha+Beta	CoiledCoil	Small	Multidomain	Others
C	0.00	0.59	7.75	-7.77	-4.25	-7.77	347.31	-9.49	-38.30
W	0.00	-12.17	25.36	-10.92	-4.66	-10.92	8.74	-11.32	96.03
M	0.00	9.38	-10.80	5.67	5.05	5.67	-21.23	2.33	-24.90
H	0.00	-7.63	-6.90	-8.70	-2.77	-8.70	-2.88	-19.25	-38.58
Y	0.00	-2.09	11.28	-3.83	0.95	-3.83	10.63	2.98	13.84
Q	0.00	9.74	0.39	-6.72	-2.05	-6.72	-1.27	-2.99	17.30
F	0.00	-0.84	3.34	-0.49	0.48	-0.49	-2.03	-0.88	-4.08
N	0.00	-6.83	11.29	-6.44	-3.92	-6.44	5.73	-7.62	33.64
P	0.00	-10.03	7.38	0.04	-2.67	0.04	23.26	5.85	-3.68
R	0.00	5.29	-7.45	2.33	5.15	2.33	10.81	6.85	-22.88
I	0.00	-0.67	-4.23	7.22	4.74	7.22	-17.06	7.35	-4.08
T	0.00	-9.69	12.44	-7.03	-4.76	-7.03	-3.09	-3.52	0.67
K	0.00	11.57	-1.70	1.34	4.05	1.34	5.74	10.37	-11.96
D	0.00	-0.42	0.91	-1.52	-0.75	-1.52	-10.77	-1.45	-16.14
S	0.00	-5.71	12.59	-8.60	-5.15	-8.60	10.05	-14.92	14.27
E	0.00	15.98	-7.80	4.21	10.27	4.21	3.45	11.79	-14.06
V	0.00	-9.45	5.28	5.50	1.91	5.50	-17.59	0.83	2.81
G	0.00	-15.37	5.04	0.84	-3.77	0.84	-3.69	-3.32	5.76
A	0.00	6.17	-15.80	6.14	-3.03	6.14	-27.68	0.58	9.16
L	0.00	11.35	-12.17	2.06	0.19	2.06	-17.30	1.04	2.17

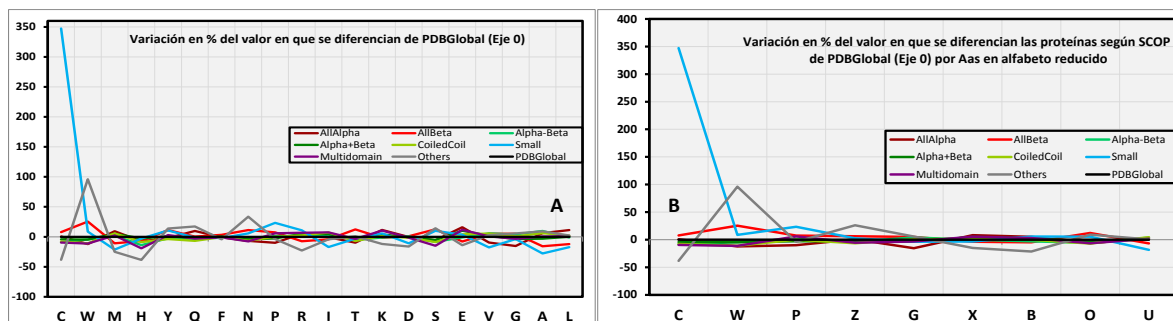


Figura 3.1.1.3. Frecuencias relativas de las proteínas agrupadas por composición de estructura secundaria según SCOP.

En la figura 3.1.1.3., también es evidente que la Cisteína está casi la mitad de las veces en proteínas clasificadas como “otras”, que en las demás proteínas de la clasificación SCOP, lo que sugiere que estas proteínas se ven más estabilizadas por fuerzas no covalentes que la demás proteínas. En este grupo de proteínas clasificadas como “otras”, también es llamativo que el Triptófano sea un aminoácido mucho más frecuente que en las demás clasificaciones proteicas,

con una frecuencia de aparición, un 96% mayor que en las proteínas de PDBGlobal, que puede encontrar explicación en el hecho de que algunas proteínas pequeñas contienen un motivo estructural conocido como “Cremallera de triptófano” (“Tryptophan zipper”) (Cochran et al, 2001; Snow et al, 2004; Du et al, 2004)

En estas proteínas, la Histidina, aparece un 38.7% menos que en el grupo de referencia, al igual que los otros dos aminoácidos básicos Arginina (22.8% menos frecuente) y Lisina (11.96% menos frecuente), lo que sugiere que estas proteínas pueden aparecer más frecuentemente en ambientes cargados positivamente, o interactuando con moléculas de carga negativa.

En la figura anterior, la escala de las diferencias por encima de las frecuencias del PDBGlobal, se ajustaron a más de un 300% para mostrar las diferencias en Cisteína de las proteínas “small”. Con ello, las diferencias de las demás proteínas se hicieron visualmente menos evidentes. En esta figura, a una escala en las abscisas, de entre -100 y 100, correspondiente a las diferencias normalizadas, salta a la vista que el Triptófano también presenta una diferencia significativamente alta en las proteínas agrupadas como “otras. Esta diferencia se comentó en páginas anteriores.

3.1.2. Frecuencias de aminoácidos de las proteínas clasificadas por su localización celular.

Se presentan las frecuencias de los aminoácidos en proteína agrupadas por su localización celular. Se hizo una prueba de t-student pareada, comparando cada grupo de proteínas con la generalidad de las proteínas representada en el PDB, y dio como resultado que las frecuencias de cada uno de los grupos de proteínas, es diferente en su frecuencia de aminoácidos, con la generalidad de las proteínas, representada en el PDBGlobal.

Tabla 3.1.2.1. Frecuencias de aminoácidos en proteínas agrupadas según su localización celular.

Aa	PDBSample	Citoplasma	MembrCel	MatrExtrCel	MatrMit	Citoesqlt	Organelas	Golgi	MembrMit
C	1.26	1.34	1.28	2.46	1.40	1.56	1.74	1.66	1.23
W	1.37	1.11	1.38	1.69	1.05	1.19	1.28	1.47	1.57
M	2.34	2.48	2.42	1.90	2.61	2.60	2.46	2.61	2.36
H	2.81	2.71	2.48	3.18	2.43	2.40	2.74	2.79	2.86
Y	3.41	3.15	3.47	3.61	3.02	3.27	3.25	3.85	3.37
Q	3.62	3.58	3.80	3.98	3.96	4.16	4.05	4.23	3.22
F	3.89	3.98	4.13	4.54	3.94	4.10	4.08	4.61	4.66
N	4.23	3.94	4.16	4.33	3.86	3.97	4.12	4.22	4.54
P	4.64	4.55	4.54	5.81	4.99	4.15	4.91	4.87	5.22
R	4.91	4.98	4.84	5.01	4.70	5.11	4.92	4.80	4.14
I	5.54	5.82	5.50	4.53	5.79	5.70	5.41	5.07	5.48
T	5.57	5.26	5.68	5.36	5.20	5.37	5.30	5.25	5.41
K	5.66	5.95	5.59	5.13	6.19	6.80	6.35	6.14	6.22
D	5.86	5.65	5.74	5.56	5.22	6.02	5.49	5.82	5.10
S	6.06	5.85	6.29	6.93	6.02	6.27	6.43	6.52	6.42
E	6.52	6.96	6.35	5.42	6.56	7.55	6.83	6.89	5.90
V	7.06	7.22	7.11	6.68	7.06	6.74	6.73	6.66	7.05
G	7.87	7.64	7.66	8.15	8.19	6.82	7.19	6.47	7.74
A	8.53	8.52	8.48	7.42	8.59	7.08	7.37	6.57	8.42
L	8.82	9.31	9.08	8.31	9.19	9.15	9.36	9.51	9.08

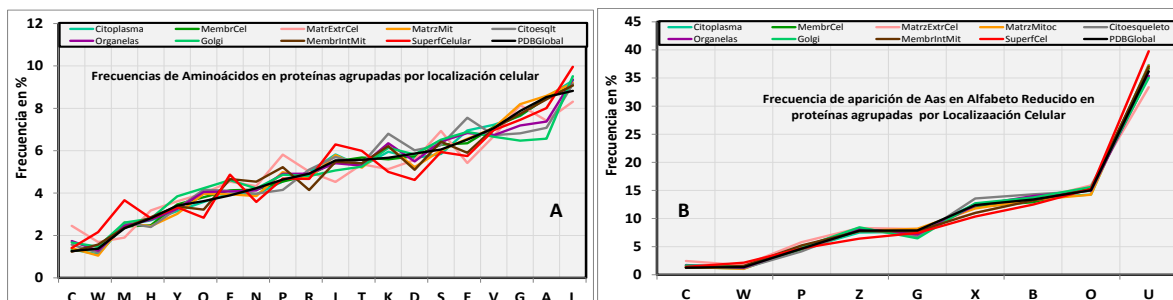


Figura 3.1.2.1. Frecuencias de aminoácidos individuales en las proteínas agrupadas por su localización celular, comparadas con las frecuencias de la muestra PDBGlobal, y ordenadas ascendentemente.

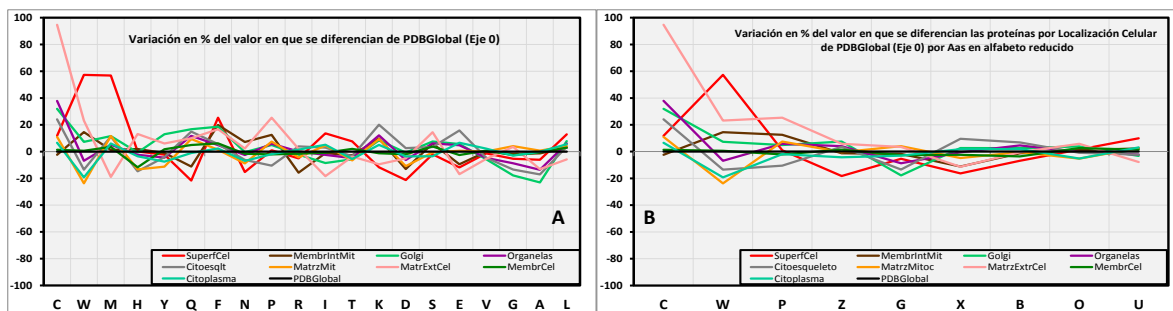


Figura 2.1.2.2. Frecuencias relativas de aminoácidos en las proteínas agrupadas por su localización celular.

Tendencias generales y Diferencias importantes

Como puede observarse en la tabla 3.1.2.1., las diferencias entre las frecuencias de proteínas agrupadas según su localización celular y las del PDBGlobal, son mucho menores que las que hay entre las proteínas agrupadas por composición de estructura secundaria y el PDBGlobal.

En la tabla 3.1.2.1 y la figura 3.1.2.1., es evidente que la mayor diferencia se presenta en las proteínas de la Superficie celular, seguido por diferencias puntuales en proteínas de la matrix extracelular y del complejo de Golgi.

En las proteínas de la superficie celular, hay un mayor contenido de los residuos hidrofóbicos Triptófano, Metionina, Fenilalanina, Isoleucina, y Leucina, y menor contenido de algunos aminoácidos polares como Arginina, Glutamina, Lisina, Glutámico, Serina y Aspártico. Esto es coherente con el hecho de que las proteínas de la superficie celular interactúan con superficies hidrofóbicas, tanto con la membrana celular, como con proteínas de membrana (Yamada, 1983, Ferguson y Williams, 1988; Doer, 2009).

Para otros grupos de proteínas, la Glicina y la Alanina son menos frecuentes en las proteínas del Complejo de Golgi que en las demás proteínas. En las proteínas de la Matriz extracelular, el aminoácido menos frecuente es el Triptófano, lo cual es de esperar, dado que es uno de los aminoácidos más hidrofóbicos y a su vez el más grande, lo que puede hacer a las proteínas menos solubles en un medio esencialmente acuoso como este medio extracelular. Por otro lado, en estas

proteínas, el aminoácido Cisteína es casi dos veces más frecuente que en las demás proteínas, lo que lleva a pensar que estas proteínas se estabilicen más frecuentemente mediante enlaces de disulfuro, que por fuerzas débiles, y/o cumpliendo una función nucleófila. Es de resaltar que el medio extracelular es un ambiente más oxidativo que el citoplasma, por lo que los puentes de disulfuro se ven favorecidos en la matrix extracelular, lo que no ocurre en el interior citoplasmático celular (Sevier y Kaiser, 2002; Oka y Bulleid, 2013).

Las proteínas de la membrana interna mitocondrial, presentan pequeñas diferencias en las frecuencias de aparición de sus aminoácidos individuales, con menores frecuencias que PDBGlobal en Glutamina (Q), Arg, Glutámico (E), y Aspártico (D) y mayores frecuencias que PDBGlobal en Fenilalanina, Asparragina, Prolina, y Lisina.

Mientras tanto, las proteínas de la matriz extracelular, Metionina (M), Isoleucina (I), Lys y Glutámico (E) están en menor frecuencia que PDBGlobal, y Prolina (P), está en mayor frecuencia. Que la Prolina esté en mayor frecuencia en estos aminoácidos se puede explicar en el hecho de que el Colágeno, rico en prolina, es la proteína más abundante de la Matriz extracelular. La Glicina (G) también está ligeramente más abundante en las proteínas de esta localización celular, lo que refuerza esta observación.

En todos los grupos de proteínas, la Leucina es el aminoácido más frecuentemente encontrado. En las proteínas de la matriz extracelular, la Cisteína presenta una frecuencia de casi el doble en comparación con las proteínas de PDBGlobal. La matriz extracelular es un ambiente diferente de el del citoplasma y es parcialmente oxidante, con lo que las proteínas de este espacio, pueden tener más puentes de disulfuro que otras proteínas que “viven” en ambientes más reductores.

Las proteínas encontradas en el complejo de Golgi, presentan frecuencias menores en los aminoácidos Alanina (A) y Glicina (G). Las de citoesqueleto y organelas, presentan frecuencias menores de Alanina (A). Las proteínas de la matriz extracelular, presentan frecuencias más altas de Prolina (P), y más bajas de Glutámico (E) y de Alanina (A).

Al agrupar los aminoácidos en un alfabeto reducido, el comportamiento de los residuos en todas las proteínas agrupadas por localización celular, se asemejan mucho más al PDBGlobal.

Aunque no es tan notorio en la figura 3.1.2.1., al normalizar las diferencias entre grupos de proteínas y el PDBGlobal, salta a la vista que las proteínas de la matriz extracelular también presentan diferencias importantes con el PDBGlobal. La diferencia más conspicua es la de la frecuencia de Cisteína, de casi 2 veces más (cerca del 100%), con respecto al PDBGlobal, en las proteínas de matriz extracelular, lo que sugiere que estas proteínas son más ricas en puentes de disulfuro ó usan las Cisteínas como nucleófilos, ayudadas por el ambiente más oxidativo que otras localizaciones celulares. Las proteínas de la matriz extracelular, presentan en contraste, un menor contenido de Metionina que todas las demás proteínas clasificadas por localización celular y que

las proteínas de la muestra del PDB. También presentan menores contenidos de Isoleucina, y de ácido Glutámico que las demás proteínas.

También es notorio que presentan alrededor de un 60% más de Triptófanos e Histidinas en las proteínas de la superficie celular, que las proteínas de PDBGlobal. En las proteínas de superficie celular se pueden observar frecuencias menores de los aminoácidos Glutamina, y Asparragina, Glutámico y Aspártico.

Las proteínas encontradas en el complejo de Golgi presentan frecuencias menores en los aminoácidos Alanina y Glicina. Las de citoesqueleto y organelas, presentan frecuencias menores de Alanina (A).

Las menores diferencias entre todos los grupos de proteínas y el PDBGlobal, se dan en los residuos hidrofóbicos y los residuos con cadenas laterales hidroxílicas.

3.3.1.3. Frecuencias de aminoácidos de las proteínas clasificadas por el organismo del que se aislaron.

Se seleccionaron proteínas de una variada gama de trece organismos, representantes tanto de Eucariota y Procariota, como de Archaea. Los resultados, se presentan a continuación.

Se realizó una prueba de t-student pareada para comparar cada grupo de proteínas, y se encontró que las frecuencias son diferentes.

Tabla 3.1.3.1. Frecuencias de aparición de los veinte aminoácidos estándar en las proteínas agrupadas por el organismo del que se aislaron.

Aa	PDBSample	Human	Mouse	Rattus	Bos	Drosophi	Arabidpsis	Saccharom	Pseudom	Bsubtilis	Ecoli	Thermotoga	Thermus	Mycobact
C	1.26	2.01	2.20	1.86	2.27	1.59	1.42	1.12	0.96	0.75	1.03	0.70	0.32	0.71
W	1.37	1.38	1.59	1.19	1.26	1.32	1.21	1.08	1.33	0.96	1.32	0.92	1.29	1.12
M	2.34	2.36	2.05	2.41	2.31	2.44	2.22	2.03	1.95	2.65	2.61	2.55	1.84	2.17
H	2.81	2.97	2.66	2.66	2.59	3.11	2.47	2.55	3.11	3.26	2.74	3.04	2.17	3.16
Y	3.41	3.39	3.69	3.34	3.25	3.45	2.93	3.53	3.05	3.39	3.06	3.19	2.84	2.37
Q	3.62	4.27	4.42	4.24	3.69	4.29	3.15	3.69	4.45	3.70	4.23	1.65	2.23	3.27
F	3.89	4.15	4.03	4.27	4.57	4.14	4.18	4.31	3.26	3.87	3.55	4.47	3.59	2.82
N	4.23	3.99	4.08	4.06	4.14	4.51	4.05	5.18	3.19	4.18	4.24	3.64	1.79	2.55
P	4.64	4.85	4.99	4.76	5.19	4.47	4.70	4.19	4.96	3.88	4.43	3.96	6.22	5.02
R	4.91	5.05	4.65	5.19	4.94	5.49	4.63	4.16	6.57	4.18	5.21	5.49	8.01	6.88
I	5.54	5.05	4.43	5.17	5.31	5.34	5.52	6.60	4.43	6.75	5.61	6.97	2.86	4.35
T	5.57	5.29	6.17	5.41	5.76	4.98	5.28	5.38	4.90	5.51	5.43	4.37	4.07	5.78
K	5.66	6.08	5.94	6.19	6.12	5.98	6.38	7.35	3.54	6.64	5.00	7.75	4.17	2.54
D	5.86	5.36	5.24	5.46	5.16	6.09	5.94	6.23	5.77	5.72	5.96	5.37	3.83	6.64
S	6.06	7.01	8.60	6.73	6.45	6.35	7.33	7.09	5.91	6.02	5.34	5.33	3.21	5.55
E	6.52	6.80	6.07	6.90	6.43	6.99	6.75	6.94	6.30	7.66	6.60	9.48	9.16	5.98
V	7.06	6.80	6.75	6.84	6.73	6.84	7.58	6.48	6.79	6.73	7.10	8.58	8.62	8.26
G	7.87	6.96	7.07	7.06	7.46	6.21	7.54	6.16	8.71	7.32	7.77	7.34	9.22	8.65
A	8.53	6.00	6.00	7.00	7.00	7.00	7.00	6.00	10.43	8.00	9.00	6.00	11.00	12.00
L	8.82	9.44	9.06	9.18	9.02	9.13	9.39	9.44	10.39	8.69	9.21	9.13	12.70	9.92

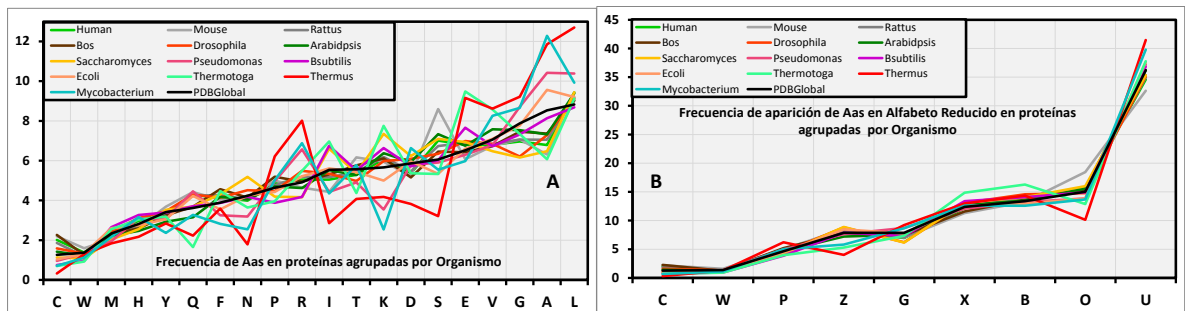


Figura 3.1.3.1. Frecuencias de aminoácidos individuales en proteínas agrupadas por el organismo del que se aislaron.

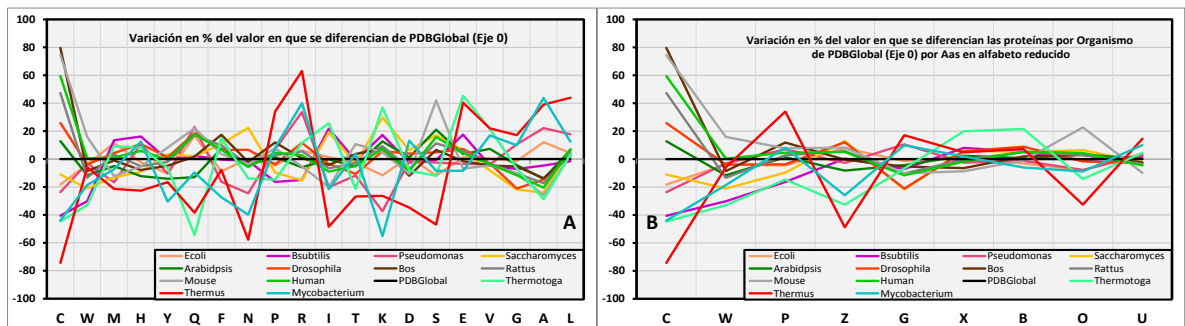


Figura 3.1.3.2. Frecuencias relativas (normalizadas) de aminoácidos individuales en proteínas agrupadas por el organismo del que se aislaron.

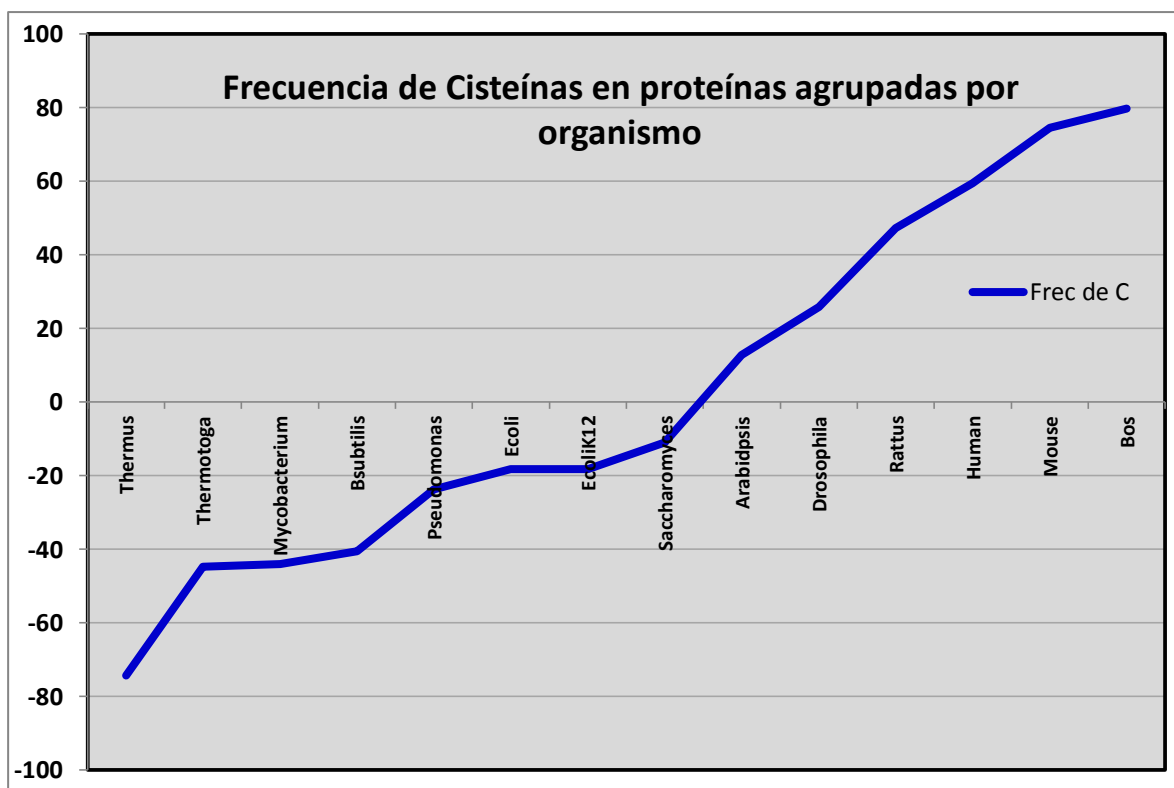


Figura 3.1.3.1.1. Contenido de Cisteína con respecto al PDBGlobal, en proteínas agrupadas por el organismo en que se encontraron.

Tendencias generales y Diferencias importantes

Como puede observarse tanto en la tabla 3.1.3.1., en la figura 3.1.3.1. y en la figura 3.1.3.1.1., las diferencias entre las frecuencias de proteínas agrupadas según el organismo de origen, son mayores que las que hay entre las proteínas agrupadas por composición de estructura secundaria y el PDBGlobal. Además, en la figura 3.1.3.1.1. es claro que los organismos pertenecientes a Achaea y a los Procariota más antiguos, presentan los menores contenidos de Cisteína, mientras que los organismos eucariotes, tienen mayor contenido de este aminoácido y este contenido aumenta en la medida en que el organismo es evolutivamente más reciente. Con esto, consideramos que el contenido de Cisteína en las proteínas de un organismo, sirve como indicador de la antigüedad de ese organismo. Esto es algo que según nuestra búsqueda bibliográfica, no se había explorado antes, aunque sí se había reportado a la Cisteína como indicador de la complejidad de un organismo (Miseta y Csutora, 2000). En la figura 3.1.3.1. Se observan diferencias, principalmente, en los residuos de Cisteína, Triptófano, Prolina, los de cadena lateral amida y Glicina en todas las proteínas seleccionadas.

Los residuos de Cisteína presentan la mayor variabilidad en todas estas proteínas, lo que lleva a inferir que las proteínas presentan diferentes grados de estabilización de la estructura 3D, soportándose en algunos casos en la formación de puentes de disulfuro, mientras que en otros

casos, se apoya en la formación de interacciones débiles. En la figura es evidente que las mayores diferencias están en las proteínas pertenecientes a los organismos *Thermus*, *Thermotoga* y *Mycobacterium*, que además, son los organismos evolutivamente más antiguos de entre los organismos seleccionados. Distancias evolutivas en comparación con *Homo sapiens*: (*Thermus thermophilus* y *Thermotoga maritima*: 2535.8 Millones de años (My por su sigla en inglés); *Thermus* y *Thermotoga*, son bacterias extremófilas que viven en ambientes extremos de alta temperatura, mientras que *Mycobacterium* es una bacteria aerobia que se caracteriza por ser muy ácido-resistente. Es decir, las tres bacterias pueden vivir en ambientes extremos, por lo que sus proteínas también. (Huber et al, 1986; Oshima e Imahori, 1974; Vandai, 2009). En su orden, la antigüedad de los organismos de la gráfica, son: *Saccharomyces* (1215.8 My), *Arabidopsis*, 1369.0 My; *Drosophila*, 782.7 My; *Rattus*, 92.3 My; lo anterior implica que el contenido de Cisteína tiene una alta relación con la antigüedad evolutiva. Sugerimos ampliar este trabajo, incluyendo un mayor número de proteínas, para lo cual, es preferible usar la base de datos de SwissProt, que al contener secuencias proteicas y no estructuras, cuenta con un mucho mayor número de proteínas almacenadas.

En las figuras 3.1.3.1. y 3.1.3.2., se puede observar que las proteínas de las gram negativas, *Thermus thermophilus*, y *Thermotoga maritima* cuyas temperaturas óptimas de crecimiento están alrededor de los 65°C, y 80°C, respectivamente (Beffa et al, 1996; Huber et al, 1986), presentan los menores contenidos de Cisteína (C), hasta un 70% menos que en las demás proteínas; Glutamina (Q), hasta un 58% menos, y Tirosina (Y), con 27% menos (en cuanto a este último aminoácido, también *Mycobacterium* presenta valores más bajos que en las demás proteínas).

Estos menores contenidos de Cisteína (un 70% menos de este residuo), fueron las proteínas de *Thermus*, una bacteria termófila. Esta bacteria se aisló originalmente de fumarolas hidrotermales, que contienen significativas cantidades de ácido sulfhídrico (H₂S), y viven en un ambiente reductor, que haría casi imposible la formación de puentes de disulfuro, por lo que posiblemente las Cisteínas son innecesarias y casi inútiles para sus proteínas, en términos de la estabilización de la estructura mediante la formación de enlaces de disulfuro. De allí, que este aminoácido puede haber sido seleccionado negativamente en los genes de estas proteínas.

En cuanto a los residuos de Cisteína, *Thermotoga* también presenta un menor contenido de este residuo (un 45% menos de lo que contiene PDBGlobal), explicado posiblemente en el hecho de que las aguas termales en que viven, también pueden contener cantidades relativamente altas de ácido sulfhídrico (H₂S), y por tanto, sería válido ofrecer la misma explicación que se sugiere para *Thermus*.

En cuanto a las proteínas de los demás organismos, salta a la vista que los organismos eucariotes presentan mayores contenidos de Cisteína en sus proteínas, y los procariotes presentan menores contenidos de este aminoácido. La mayoría de las Cisteínas en proteínas, aparece formando puentes de disulfuro, y esta formación de enlaces representa la función más importante de las Cisteínas en las proteínas (Neves et al, 1999), por lo que consideramos la frecuencia de aparición

de este aminoácido en proteínas, desde el punto de vista de su aporte a la estabilidad proteica estructural mediante el aporte de puentes de disulfuro.

El mayor contenido de Cisteína en eucariotes, ocurre probablemente debido al ambiente oxidante que produce el metabolismo oxidativo más activo en procariotes, que en eucariotes y archaea. Esto puede producir microambientes suficientemente oxidantes como para que ocurran los enlaces de disulfuro. Los organismos procariotes y la levadura como el organismo eucariote más antiguo de nuestra selección de organismos, presentan niveles más bajos de Cisteína que el PDBGlobal y que los organismos eucariotes. El contenido gradualmente más alto de Cisteínas en proteínas de organismos más recientes, ya se había observado en estudios anteriores (Brooks y Frescos, 2002). De acuerdo con esto, consideramos que probablemente la Cisteína, pueda usarse como una especie de indicador de antigüedad evolutiva de una proteína, ó de un organismo, si hablamos del contenido de Cisteína del proteoma de un organismo.

También hay particularidades en las frecuencias de aminoácidos de *Thermus*, en comparación con las de *Thermotoga*. *Thermus* presenta los menores contenidos de Alanina (A), Isoleucina (I), Treonina (T), Ácido aspártico (D) y Serina (S), y los mayores contenidos de Prolina (P), Arginina (R) y Leucina (L). *Thermotoga*, por su parte, presenta los menores contenidos de Alanina (A) y los mayores contenidos de Glicina (G). Aunque ambas bacterias viven en ambientes acuosos a altas temperaturas, *Thermus* ha sido aislada en fumarolas marítimas en las profundidades marinas, con lo que está soportando presiones mucho mayores, que *Thermotoga*, que se halló en lagunas termales, y esto puede tener influencia en la composición de aminoácidos. También es importante resaltar que la diferencia de temperatura óptima de crecimiento entre ambas bacterias (*Thermus*: 65°C; *Thermotoga*: 80°C), puede ser importante en la composición de aminoácidos de sus proteínas termoestables.

Los contenidos de Ácido Glutámico y Valina, en las bacterias termófilas son más altos que las de las proteínas de los demás grupos, incluyendo el PDBGlobal, lo que sugiere que estos residuos podrían participar en conferir más alta estabilidad térmica a una estructura proteica.

El aminoácido más común en las proteínas de *Thermus* es Prolina, que en el caso del colágeno, es responsable por la termoestabilidad de esa proteína, en su forma hidroxilada (Gustavson, 1955; Hall y Reed, 1957; Jimenez y Rosenbloom, 1974; Zscherp et al, 2003), lo que lleva a pensar que este residuo de aminoácido esté aportando termoestabilidad a las proteínas de *Thermus*.

También es llamativo que en las proteínas de *Thermus*, hay un menor contenido de los aminoácidos con cadenas laterales amida (Z: Asparragina y Glutamina), alrededor de un 50% menos. Esto podría estar parcialmente explicado en dos hechos: primero, las amidas de las cadenas laterales de estos residuos, forman enlaces de H, que no serían muy estables a altas temperaturas, y por tanto no aportarían a la estabilización de la estructura 3D de las proteínas de este organismo (Shimoni y Glusker, 1995).

Segundo, las Asparraginas y Glutaminas podrían conferir menor estabilidad estructural a las proteínas y a que las proteínas de *Thermus* deben ser proteínas muy estables en su estructura. Los residuos de Asparragina (N) y Glutamina (Q) se han identificado como especialmente altos en proteínas tendientes a formar cuerpos amiloides autopropagantes, como en los priones. Estos cuerpos amiloides corresponden a proteínas que han cambiado su plegamiento normal en las células (Michelitsch y Weissman, 2000).

Las proteínas de *Thermotoga* presentan niveles elevados en los residuos de aminoácidos ácidos y básicos, con contenidos un 20% más altos en estos residuos. Esto podría contribuir a la formación de un corazón en las proteínas globulares de *Thermotoga*, formado por aminoácidos que se atraen mutuamente, más que por efecto de la hidrofobicidad en un corazón hidrofóbico. Los contenidos de los residuos con cadenas laterales hidroxilas, también están en menor proporción en las proteínas de *Thermotoga*, que en la muestra PDBGlobal.

Hay menor contenido de los residuos de Asparragina y de Glutamina en general, en organismos procariotes, y un poco más en eucariotes. Esto podría estar parcialmente explicado en el hecho de que estos aminoácidos participan en las modificaciones postraduccionales de proteínas eucariotes, inexistentes en las de eucariotes.

Mycobacterium presenta sus particularidades también: tiene los menores contenidos de Tirosina (Y) y Lisina (K), y los mayores contenidos de Alanina (A), que pueden estar implicados en la resistencia estructural de algunas proteínas, a ambientes ácidos.

Tal vez una de las cosas más interesantes de la comparación de los contenidos de aminoácidos, es la de identificar residuos que eventualmente contribuyan a conferir propiedades particulares a las proteínas, para ser usados en el rediseño de estas moléculas.

3.1.4. Frecuencias de aminoácidos de las proteínas clasificadas por Phylum.

Se presentan y discuten los resultados de las frecuencias de aparición de aminoácidos en las proteínas de organismos agrupados por Phylum. También se presentan los resultados de las pruebas pareadas de t-student, que muestran que las frecuencias entre estas proteínas y la generalidad de las proteínas en el PDBGlobal, son diferentes.

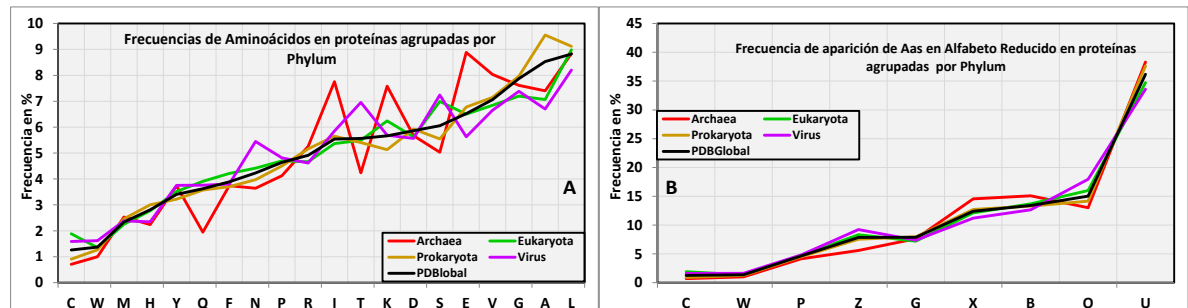


Figura 3.1.4.1. Frecuencias de aminoácidos en proteínas agrupadas por Phylum, y comparadas con PDBGlobal.

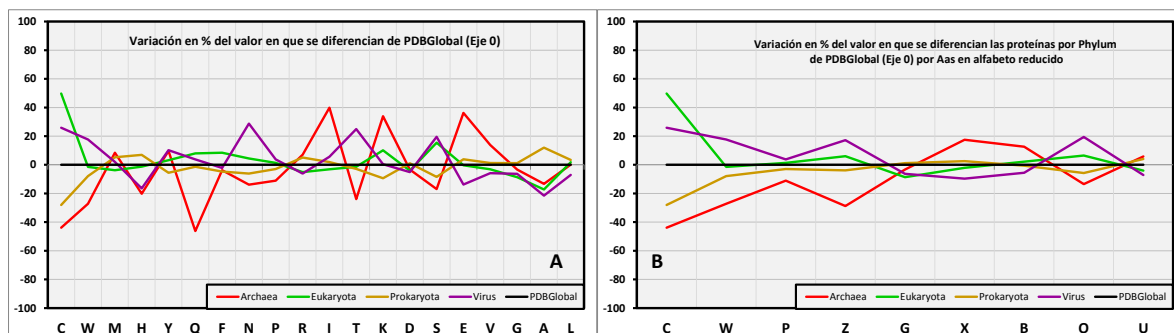


Figura 3.1.4.2. Frecuencias normalizadas de aminoácidos en proteínas agrupadas por Phylum, y comparadas con PDBGlobal.

Tendencias generales y Diferencias importantes

En las figuras 2.1.4.1. y 2.1.4.2., se observan las siguientes cosas en común: en la parte B de la figura 3.1.4.1., se ven pocas diferencias entre las frecuencias de las proteínas agrupadas por Phylum, con la excepción más notoria, de las proteínas de Archaea. En la parte A, de la misma figura, se observan también algunas diferencias en las proteínas virales. En mucha menor medida, se ven algunas diferencias en las frecuencias de Procariota. Las proteínas de organismos eucariotes presentan menos diferencias con el PDBGlobal. Más de la mitad de las proteínas del PDB, corresponde a proteínas eucariotas. Esto puede explicar la poca diferencia entre estas y el PDBGlobal. Las procariotas constituyen alrededor del 40%. También, además del contenido de proteínas eucariotes en la totalidad del PDB, las muestras de cada grupo de proteína, fueron de diferente tamaño: 10.361 proteínas eucariotes, contra 7.824 de procariotes, 1.120 de Archaea y 759 de virus).

Por otro lado, los residuos de Metionina, Tirosina, Fenilalanina, Arginina, Aspártico, Glutámico y Leucina, son los que presentan menos diferencias entre las proteínas de diferentes Phyla. Es decir, las proteínas de los diferentes Phyla, presentan contenidos similares de estos organismos. Esto sugiere que estos residuos pueden ser determinantes para el andamiaje de las proteínas y que son de aparición temprana en la historia evolutiva.

Los residuos de Cisteína, tal como se comentó en la sección anterior, 2.1.3. aparece en mayor cantidad que el PDBGlobal, en proteínas de Eucariotes, y en menor proporción, en proteínas de Procariota y Archaea, lo que es fácilmente explicable por el mayor metabolismo oxidativo en Eucariotes, y por tanto la generación de microambientes celulares más oxidantes, adecuados para la formación y mantenimiento de puentes de disulfuro, y además que los eucariotes presentan mayor cantidad de proteínas multiméricas cuyas subunidades se unen mediante puentes de disulfuro, que los demás Phyla.

Que las proteínas de eucariotas, se encuentran más expuestas a ambientes oxidantes, que las proteínas de Archaea y Procariota, puede entenderse evolutivamente en que tanto Archaea como Procariota, viven en el planeta desde que había una atmósfera reductora, mientras que la mayoría de Eucariotes sólo apareció después de que la atmósfera se hizo más rica en oxígeno y por tanto más oxidante, lo que facilita la formación de puentes de disulfuro. Eso lleva a pensar que la formación de puentes de disulfuro por las Cisteínas puede haber sido un evento de aparición más tardía en la historia evolutiva de las proteínas y los organismos.

De manera curiosa, los virus también presentan un contenido mayor de Cisteínas, que las proteínas del PDBGlobal y las Procariotes y Archaea, que puede encontrar explicación en que mucha de las proteínas de cápsides virales se estabilizan y se organizan multiméricamente mediante enlaces de disulfuro (Regeimbal y Bardwell, 2002; Messens y Collet, 2006; Berkmen, 2012).

En Archaea, las proteínas presentan mayores frecuencias en los residuos de Isoleucina, Lisina, y Glutámico, con diferencias de entre el 30 y el 40%, y menores frecuencias en los residuos de Cisteína y Asparragina, con diferencias por arriba del 40%, y en menor medida, en los residuos de Cisteína, Histidina, Treonina y Serina, con diferencias cercanas al 20%.

Otra de las observaciones llamativas, es la menor frecuencia de aminoácidos con grupos Amida (Asparragina (N) y Glutamina (Q), aquí representados con la letra Z) en Eucariotes (alrededor de 30% menos); y la mayor frecuencia de los mismos residuos de aminoácidos con grupos Amida en proteínas de virus (aproximadamente un 20% más).

También es conspicuo que en algunos aminoácidos, las proteínas de virus y las de procariotes, presentan una frecuencia más o menos opuesta; es decir, en los casos en que Procariota usa un frecuentemente un aminoácido (y en frecuencia mayor que el PDBGlobal), los virus lo usan de manera menos frecuente que el PDBGlobal. A manera de ejemplo, las proteínas de procariotes son más ricas en los aminoácidos ácidos (X) y los básicos (B), que el PDBGlobal (entre un 15 y 20%), y que los virus, por el contrario, presentan menores contenidos de estos residuos ácidos y básicos que el PDBGlobal (entre un 10 y un 15% menos).

Esto también ocurre en algunos aminoácidos de Archaea y Virus; es decir, mientras las frecuencias de algunos aminoácidos en Archaea, están por encima de las frecuencias del universo proteico, en Virus, las frecuencias de los mismos residuos, están por debajo de las frecuencias de la totalidad de las proteínas almacenadas en el PDB, y viceversa.

3.1.5. Frecuencias de aminoácidos de las proteínas clasificadas por tipo de enzima

En esta sección se compararon las proteínas clasificadas por el tipo de enzima, según la Comisión de Enzimas. Aunque estas proteínas presentaron las menores diferencias es sus frecuencias con respecto a las proteína de PDBGlobal, sus frecuencias resultaron también diferentes en las pruebas pareadas de t-student.

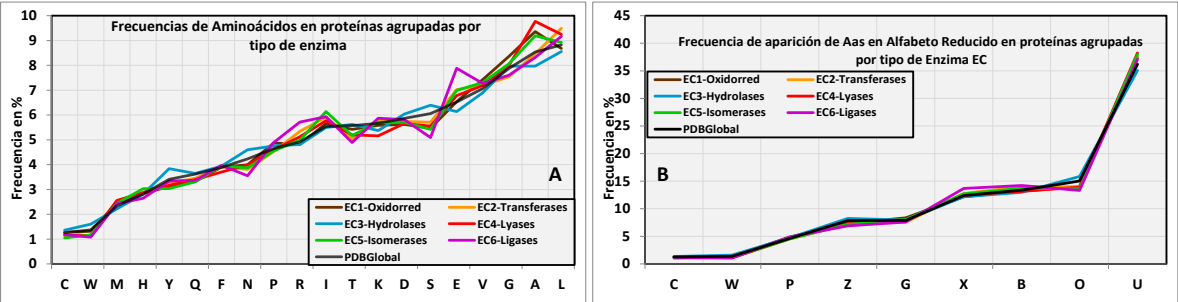


Figura 3.1.5.1. Frecuencias de aparición de los veinte aminoácidos en una muestra de proteínas enzimas, clasificadas según la Comisión de Enzimas.

Tabla 3.1.5.1. Frecuencias de aparición de los veinte aminoácidos en una muestra de proteínas del PDB clasificadas por tipo de enzima EC.

SwissProtMar6-2013	PDB muestra global	EC1-Oxidored	EC2-Transferases	EC3-Hydrolases	EC4-Lyases	EC5-Isomerases	EC6-Ligases	Promedio
1.08	1.37	1.32	1.11	1.60	1.14	1.16	1.08	1.27
1.37	1.26	1.27	1.26	1.36	1.15	1.05	1.19	1.28
2.27	2.81	2.83	2.89	2.81	2.87	3.03	2.64	2.75
2.42	2.34	2.46	2.54	2.22	2.56	2.45	2.43	2.42
2.92	3.41	3.12	3.25	3.84	3.17	3.03	3.34	3.28
3.86	3.89	3.90	3.91	3.94	3.70	3.95	3.97	3.87
3.93	3.62	3.32	3.45	3.65	3.40	3.31	3.36	3.56
4.06	4.23	3.99	3.81	4.60	3.99	3.87	3.55	4.11
4.70	4.64	4.88	4.52	4.75	4.60	4.56	4.89	4.68
5.34	5.57	5.42	5.04	5.62	5.20	5.17	4.89	5.36
5.45	5.86	5.62	5.74	6.04	5.66	5.70	5.81	5.73
5.53	4.91	4.81	5.34	4.81	5.12	4.99	5.71	5.09
5.84	5.66	5.58	5.79	5.37	5.16	5.71	5.87	5.57
5.96	5.54	5.66	5.91	5.50	5.77	6.14	5.93	5.72
6.56	6.06	5.47	5.71	6.39	5.55	5.42	5.09	5.96
6.75	6.52	6.54	7.02	6.14	6.78	6.99	7.88	6.63
6.87	7.06	7.43	7.19	6.89	7.19	7.31	7.27	7.11
7.08	7.87	8.36	7.54	7.97	7.99	8.06	7.60	7.80
8.25	8.53	9.36	8.47	7.97	9.77	9.19	8.32	8.72
9.66	8.82	8.68	9.49	8.55	9.25	8.92	9.17	9.08

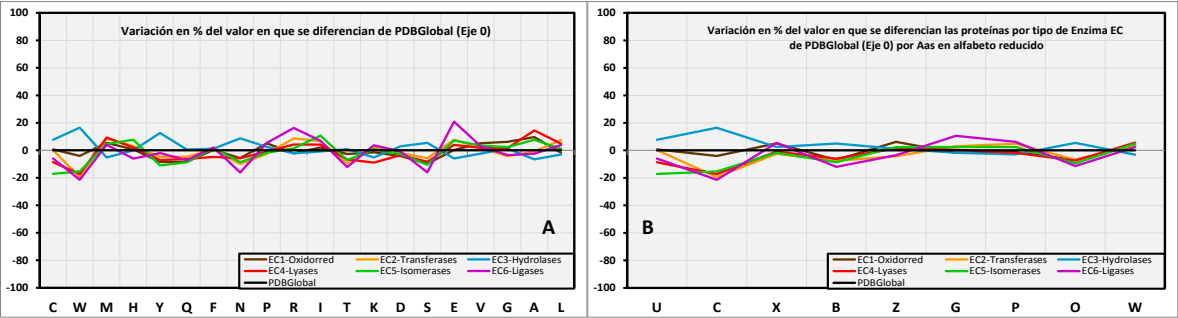


Figura 3.1.5.2. Diferencias normalizadas como porcentaje de variación de las frecuencias de aminoácidos en las enzimas con respecto a las proteínas del PDBGlobal.

Tendencias generales y Diferencias importantes

En esta figura es evidente que en esta muestra de proteínas del PDB, de enzimas clasificadas según la clasificación de la Comisión Enzimática (Webb, 1992; Moss, 1992), se dan unas frecuencias de aparición de aminoácidos, bastante cercanas a las frecuencias de la muestra de proteínas del PDB; las diferencias máximas, no superan el 20% con respecto al PDBGlobal.

Sin embargo, según las pruebas pareadas de t-student, hay diferencias significativas entre las proteínas clasificadas por tipo de enzima y el resto de las proteínas; las enzimas constituyen un grupo muy grande dentro del PDB, por lo que sus frecuencias se parecen a las de la muestra PDBGlobal. A manera de comentario, sin embargo, vale mencionar que la Alanina es un aminoácido que aparece con una frecuencia mayor en las enzimas del tipo EC-1 (Oxidorreductasas), y EC-4 (Liasas), mientras que aparece un poco menos en enzimas tipo EC-3 (Hidrolasas).

También es llamativo que el ácido glutámico aparece con mayor frecuencias en proteínas del tipo EC6-Ligasas, posiblemente como participante del sitio activo de estas enzimas, participando en la desestabilización de grupos a ser quitados o añadidos a las moléculas que actúan como sustrato, como residuo importante en fosforilaciones y defosforilaciones (Fraser et al, 2002), y como residuo capaz de ligar metales en las enzimas del tipo ligasa. (Wojciechowski y Kantrowitz, 2003). En la misma línea de pensamiento, la serina es un residuo menos frecuente en las ligasas que en los demás tipos de enzimas.

Hay una pequeña variación en aminoácidos que aparecen con menor frecuencia, pero esto no es claramente observable sin una normalización de las diferencias, con lo que generamos el gráfico 2.1.5.2. en el que se presentan las diferencias en términos de un porcentaje de variación con respecto al estándar que para este trabajo, es el conjunto de proteínas de PDBGlobal.

En la figura 3.2.5.2., se observan menores diferencias en los contenidos de residuos de aminoácidos con respecto al PDB, con las siguientes observaciones llamativas: las enzimas hidrolasas presentan una ligeramente mayor frecuencia de uso de Cisteínas que las proteínas de los demás tipos de enzimas (alrededor de un 10% más). Esto posiblemente sea debido a que las enzimas hidrolasas requieren de puentes de disulfuro para alcanzar el plegamiento correcto, y para realizar su función catalítica (Song et al, 2004; Chu et al, 2010; Kim e Ishikawa, 2013).

Otra diferencia importante, es que las enzimas hidrolasas presentan mayores contenidos de Triptófano (cerca de 20% más), que los demás tipos de enzimas, que en general, presentan menor contenido de este residuo de aminoácido que el PDBGlobal (cerca de un 20% menos).

A manera de contraste, las Isomerasas son las enzimas que usan menos Cisteínas, posiblemente debido a los tipos de reacciones que catalizan.

3.1.6. Frecuencias de aminoácidos de las proteínas agrupadas por familia proteica, según PFam.

En la presente sección, se generaron y compararon las frecuencias de aminoácidos en un conjunto de ocho familias de proteínas seleccionadas por la estabilidad de su estructura tridimensional, por su contenido de estructura secundaria o por ser familias proteicas de especial interés para otras investigaciones (ver Capítulo 2).

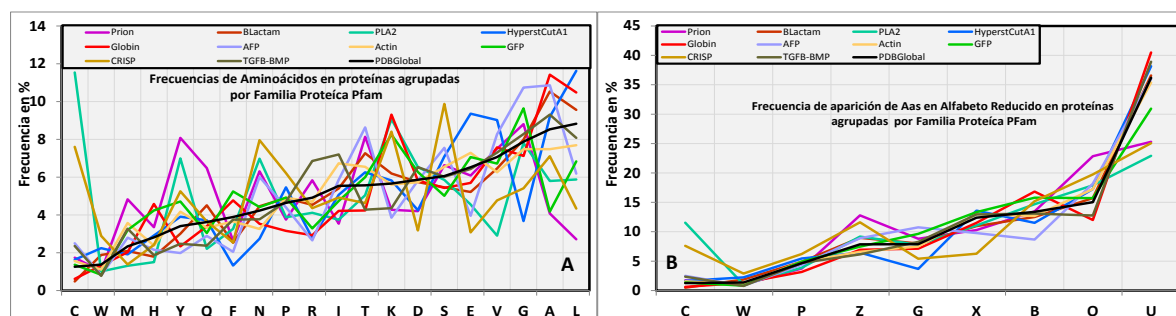


Figura 3.1.6.1. Frecuencias de los aminoácidos individuales en las proteínas agrupadas por familia proteica según PFAM y comparadas con las frecuencias de la muestra PDBGlobal.

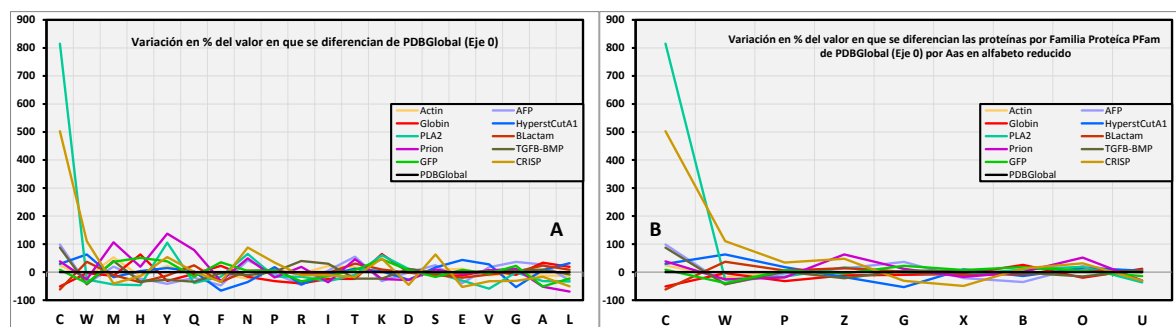


Figura 3.1.6.2. Frecuencias de los aminoácidos individuales en las proteínas agrupadas por familia proteica según PFAM, normalizadas y comparadas con las frecuencias de la muestra PDBGlobal.

Tendencias generales y Diferencias importantes

En las figuras 2.2.6.1. y 2.2.6.2., y comparando con las figuras de frecuencias de agrupaciones anteriores, es evidente que las mayores diferencias en frecuencia de aminoácidos con respecto al PEBGlobal, se presenta en las proteínas agrupadas por familia proteica, según la base de datos PFam.

De entre estas, las mayores diferencias se presentan en las proteínas Fosfolipasas A2 (PLA2), que presentan un contenido de Cisteínas especialmente alto en comparación con las proteínas del PDB (800% más de contenido).

La familia de Proteína Verde Fluorescente (GFP) es la que presenta menos diferencias importantes con respecto al PDB; sin embargo, presenta mayores frecuencias en los residuos de Aspártico,

Glutámico, Glicina, Histidina, Asparagina, y Tirosina, y menores frecuencias en Alanina, Arginina y Serina. Estas proteínas presentan características estructurales particulares, pues la proteína es un barril Beta, con dos pequeñas hélices; una dentro del barril, formando parte del cromóforo, y otra hélice muy corta, por debajo del barril (Yang et al, 1996).

En la figura 3.1.6.2., que contiene las frecuencias normalizadas (diferencias relativas), se hace muy evidente que la Cisteína está en una frecuencia anormalmente alta en las proteínas del grupo de las Fosfolipasas A2 (PLA2), lo que sugiere que estas proteínas se estabilizan muy predominantemente por enlaces de disulfuro, más que por enlaces de hidrógeno ó interacciones débiles. En algunos reportes de Fosfolipasas A2, se han encontrado numerosos puentes de disulfuro en la estructura (Beghini et al., 2000; Duan et al, 2001), llegando a veces hasta 8 puentes de disulfuro, lo que implica la presencia de al menos 16 cisteínas (Jabeen et al., 2006) en una sola proteína. En algunas Fosfolipasas A2 de venenos de serpientes se han encontrado 15 Cisteínas, lo que implica que al menos una, no es formadora de puentes de disulfuro intracatenario (Gomes-Helene, 2012).

También salta a la vista el alto contenido de Tirosinas en las proteínas de la familia de los priones, se encuentra respaldado en el hecho de que estas proteínas y los dominios formadores de priones, son ricos en este aminoácido, y también en Glutamina (King et al., 2012), que también aparece en cantidad elevada en nuestra muestra.

El Triptófano se encuentra anormalmente alto en Bacteriorrodopsinas, pues es necesario en estas proteínas para su interacción con el retinal. (Mogi et al., 1989).

3.1.7. Consolidado de Frecuencias de aminoácidos

Se hizo una comparación global de lo presentado en los puntos 2.1.1. a 2.1.6., con la idea de identificar los grupos más variables, las mayores diferencias y los aminoácidos con rangos más variables. En la figura 3.1.7., se muestran las frecuencias de los 52 grupos de proteínas del presente trabajo.

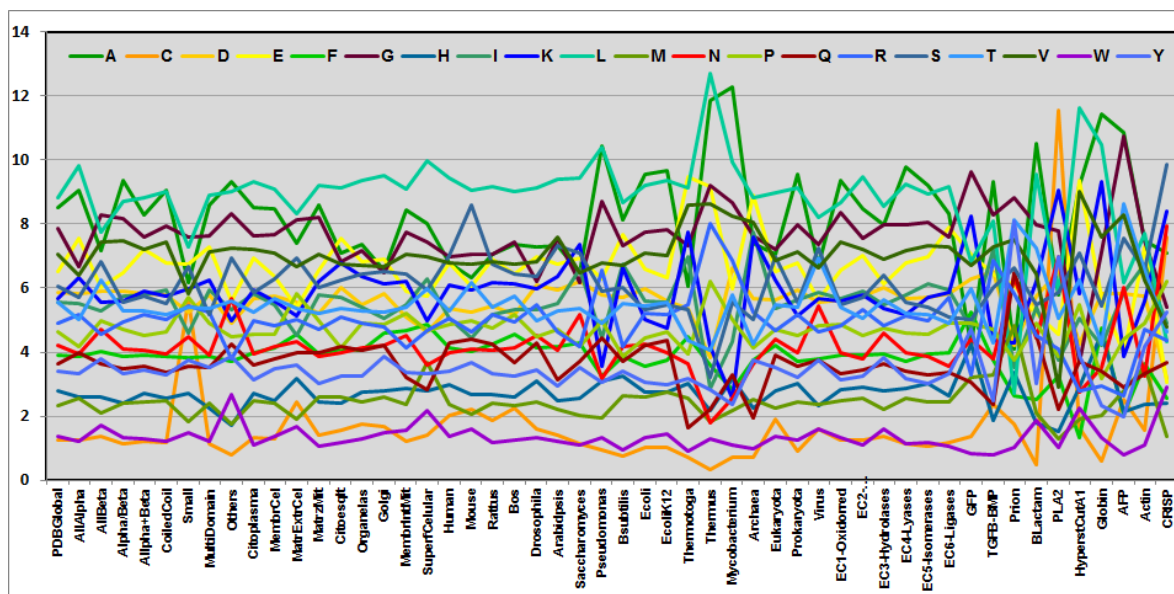


Figura 3.1.7.1. Frecuencias de los aminoácidos individuales en las proteínas en los 52 grupos de proteínas del presente trabajo.

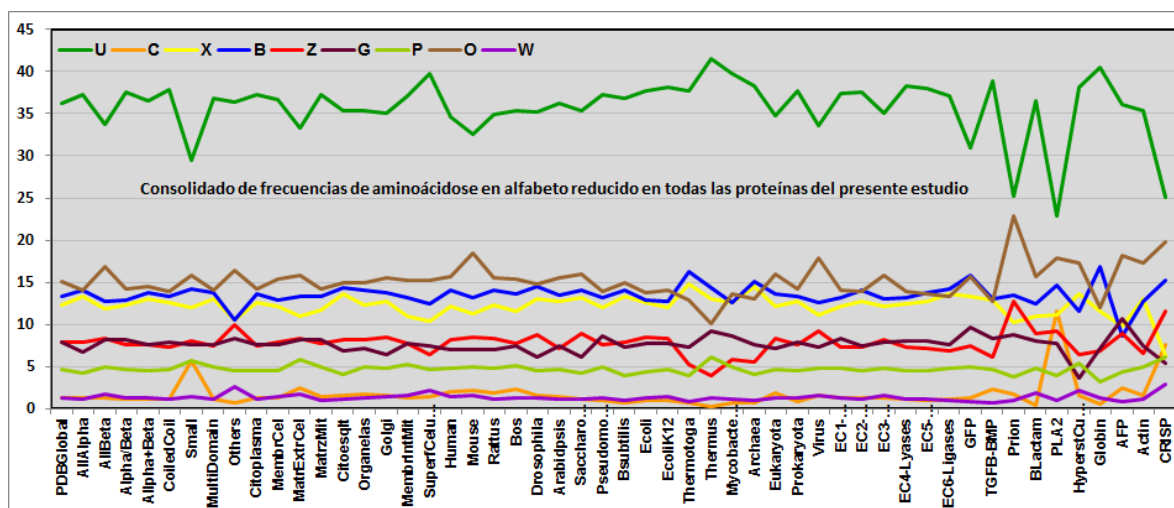


Figura 3.1.7.2. Frecuencias de las propiedades fisicoquímicas seleccionadas en un alfabeto reducido de aminoácidos, en las proteínas en los 52 grupos de proteínas del presente trabajo.

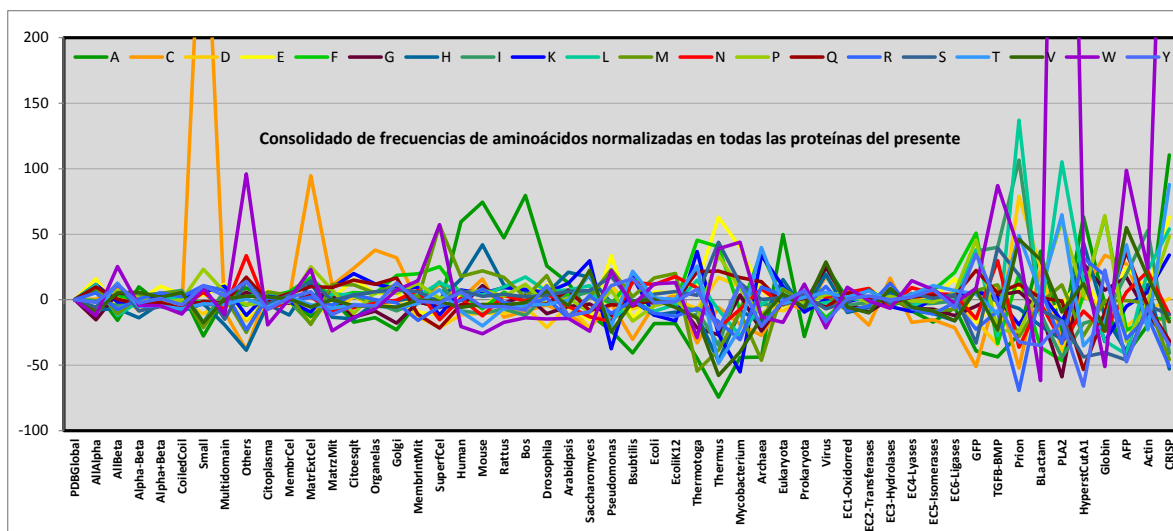


Figura 3.1.7.3. Frecuencias normalizadas de los aminoácidos individuales en alfabeto reducido, en las proteínas en los 52 grupos de proteínas del presente trabajo.

Tendencias generales y Diferencias importantes

Entre las primeras cosas que saltan a la vista, está el hecho de que las frecuencias aminoacídicas en las proteínas son variadas con respecto a la globalidad de las proteínas, representada en la muestra PDBGlobal. También es notorio que las frecuencias en las proteínas agrupadas por familia proteica, muestran claramente mayor variabilidad entre ellas, que las demás agrupaciones, seguidas por la agrupación de proteínas por Phylum. Después de ellas, el segundo grupo que presenta alta variabilidad, es el de las proteínas agrupadas por organismo. En ese mismo orden, las proteínas agrupadas por organismo, y luego las proteínas agrupadas por el contenido de estructura secundaria. Las proteínas con frecuencias más parecidas al PDBGlobal, son las agrupadas por tipo de enzima.

La Cisteína es el aminoácido que presenta la mayor variabilidad en su frecuencia de aparición de entre todos los aminoácidos, fluctuando entre 0.32% en las proteínas de Thermus, y 11.54% en las proteínas Fosfolipasas A2 (más de tres veces en comparación con el PDBGlobal). Sin embargo, la diferencia más alta presentada por un aminoácido con respecto a su aparición en la generalidad de las proteínas (PDBGlobal), es el Triptófano, que en las Fosfolipasas A2 (PLA2), está aumentado en más del 800% (Figura 3.1.7.3). El mismo aminoácido, en proteínas CRIP, está aumentado más de cinco veces con respecto al PDBGlobal.

En la figura 3.1.7.3. que contiene las frecuencias normalizadas, se hace muy evidente que la Cisteína está en una frecuencia anormalmente alta en las proteínas del grupo de las Fosfolipasas A2 (PLA2), lo que sugiere que estas proteínas se estabilizan muy predominantemente por enlaces de disulfuro, más que por enlaces de hidrógeno ó interacciones débiles. En algunos reportes de Fosfolipasas A2, se han encontrado numerosos puentes de disulfuro en la estructura (Beghini et al., 2000; Duan et al, 2001), llegando a veces hasta 8 puentes de disulfuro, lo que implica la presencia de al menos 16 cisteínas (Jabeen et al., 2006)

La Alanina y la Leucina son los aminoácidos más abundantes en proteínas, y además, presentan los rangos más amplios de variabilidad después de la Cisteína (8.20 y 9.99%, respectivamente). Alanina fluctúa entre 4.08% en proteínas del Prión, y 12.28% en proteínas de *Mycobacterium*; también es especialmente abundante en proteínas de *Thermus thermophilus*.

El Triptófano y la Cisteína, son los dos aminoácidos menos frecuentemente encontrados en proteínas. Sin embargo, es más abundante en proteínas “small”, proteínas de superficie celular, β -Lactamasas, proteínas HiperestablesA1, y proteínas ricas en Cisteína, CRISP, presentando en estas últimas, mayor abundancia que en ninguna otra proteína. El triptófano fluctúa entre las frecuencias 0.77 y 2.89, y presenta el menor rango de variabilidad en su frecuencia de aparición en las proteínas estudiadas.

La Alanina y la Leucina son los aminoácidos más abundantes en proteínas, y además, presentan los rangos más amplios de variabilidad después de la Cisteína (8.20 y 9.99%, respectivamente). Alanina fluctúa entre 4.08% en proteínas del Prión, y 12.28% en proteínas de *Mycobacterium*; también es especialmente abundante en proteínas de *Thermus thermophilus*.

La Leucina presenta la mayor abundancia de cualquier aminoácido en una proteína, al tener una frecuencia de aparición en proteínas de *Thermus*, de 12.7%. Su menor abundancia, se dan en Priones, AFP, proteínas “pequeñas” y CRISP, con frecuencias de hasta 2.71%.

La familia de Proteína Verde Fluorescente (GFP) es la que presenta menos diferencias importantes con respecto al PDB, pero presenta mayores frecuencias en los residuos de Aspártico, Glutámico, Glicina, Histidina, Asparagina, y Tirosina, y menores frecuencias en Alanina, Arginina y Serina. Estas proteínas presentan características estructurales particulares, pues la proteína es un barril Beta, con dos pequeñas hélices; una dentro del barril, formando parte del cromóforo, y otra hélice muy corta, por debajo del barril (Yang et al, 1996).

También salta a la vista el alto contenido de Tirosinas en las proteínas de la familia de los priones, se encuentra respaldado en el hecho de que estas proteínas y los dominios formadores de priones, son ricos en este aminoácido, y también en Glutamina (King et al., 2012), que también aparece en cantidad elevada en nuestra muestra.

El Triptófano se encuentra anormalmente alto en Bacteriorrodopsinas, pues es necesario en estas proteínas para su interacción con el retinal. (Mogi et al., 1989).

3.2. Frecuencias de tripletas de residuos de aminoácidos en las proteínas agrupadas por los criterios seleccionados.

Se analizan aquí las frecuencias de aparición de las tripletas de aminoácidos completas y en la forma de alfabeto reducido, como se describe a continuación.

Para el estudio de tripletas, se agruparon los 20 aminoácidos en los 9 grupos, que abarcaron las características presentadas en la tabla 3.2.1., en que se agruparon los aminoácidos hidrofóbicos A, F, I, L, M, V en la letra “U”; los aminoácidos con cadena lateral cargada positivamente H, K, R, agrupados en la letra “B”; los aminoácidos con cadena lateral hidroxílica, S, T, Y, agrupados en “O”; los de cadena lateral amida: N, Q, en “Z”; y los aminoácidos que consideramos no agrupables, por sus características particulares, C, G, W y P.

De esta manera, las posibles combinaciones de tres residuos, que con 20 aminoácidos produce $20^3 = 8000$ tripletas, se convierte en un conjunto de 9 letras que agrupan a los 20 residuos, y produce $9^3 = 729$ tripletas, que en términos de la facilidad de visualización de las figuras, es mucho más fácil de interpretar y distinguir datos interesantes.

3.2.0. Frecuencias de tripletas en proteínas de la muestra PDBGlobal

La muestra PDBGlobal abarcó 19.308 proteínas que contenían en total, 7'475.351 tripletas. A continuación se presentan los resultados del conteo de tripletas y discusiones acerca de ellos.

Sobre la aparición de tripletas de baja frecuencia en proteínas, y la probabilidad de que exista una secuencia proteica

Es importante puntualizar que en el conjunto de proteínas estudiado (PDB), se encontraron todas las posibles ocho mil (20^3 : 20 residuos en combinaciones de 3) combinaciones de tres residuos de los veinte aminoácidos estándar; esto resulta llamativo, ya que en un trabajo anterior (Lareo y Acevedo, 2005), se había reportado que con una muestra más pequeña de hélices, láminas Beta y giros, tomada del PDB, no se encontraron todas las posibles tripletas (y de manera lógica, derivado de esto, tampoco se encontrarían muchas de las posibles combinaciones de 4, 5, 6, o más aminoácidos); en una publicación nuestra, más reciente, habíamos encontrado todas las tripletas, pero algunas, como MWW, que aparecía sólo una vez, en la versión del PDB de Agosto de 2006, y se había encontrado en una enzima quitinasa de hongos (Mejia y Tischer, 2007).

Las 20 combinaciones de tres residuos, que menos aparecen en el PDBGlobal, se presentan en la tabla siguiente. Usamos estas tripletas como ejemplo, para comentar que en general, los residuos con menor frecuencia de aparición en proteínas, son los que forman las tripletas menos frecuentes. Sin embargo, esto no es absoluto. Como se puede ver en la tabla 3.3.0.1, algunas de

las tripletas menos frecuencias, las forman aminoácidos con frecuencias algo mayores, como la Histidina.

Extrapolando la información anterior, es claro que aún hay muchos millones de combinaciones de aminoácidos que no se han encontrado en proteínas, lo que puede ocurrir debido a:

1. que la evolución aún no ha “ensayado” todas las posibles combinaciones de residuos de aminoácidos para construir proteínas.
2. Que algunas combinaciones de aminoácidos han sido “ensayadas” en la evolución de estas moléculas, y se hayan “descartado”.
3. Que las combinaciones existan, pero en proteínas que no se han estudiado. Por ejemplo, las proteínas de ambientes extremos.

Tripl	TtIPDBGlobal	FrecObsTriplIPDBGlobal	FrecTriplEspPDBGlob
CMW	8	0,00000091	0,00003480
CCW	11	0,00000125	0,00001922
CHW	12	0,00000137	0,00002603
MWM	13	0,00000148	0,00011983
WCW	13	0,00000148	0,00029407
WCC	14	0,00000159	0,00022693
WWW	14	0,00000159	0,00040747
MWC	15	0,00000171	0,00009803
MWW	16	0,00000182	0,00008534
CIW	18	0,00000205	0,00001819
CMC	19	0,00000216	0,00002119
HWC	21	0,00000239	0,00005503
WCM	23	0,00000262	0,00020849
CCM	25	0,00000285	0,00002678
MWH	25	0,00000285	0,00016323
MCW	27	0,00000307	0,00008669
WWH	28	0,00000319	0,00037753
WCH	31	0,00000353	0,00028343
WPC	31	0,00000353	0,00031314
WWM	31	0,00000353	0,00026091

Tabla 3.2.0.1. Frecuencias de aparición de las 20 tripletas menos frecuentes, en las proteínas de la muestra PDBGlobal.

Es posible que ocurra una combinación de las tres posibles razones expuestas en el párrafo anterior. Sería de importancia tanto para el estudio de estructura y función proteínas como para la evolución de las mismas, conocer más sobre las combinaciones que actualmente existe, y las que posiblemente no, discriminando las que sólo aparezcan en proteínas muy antiguas que pueda llevar a pensar sobre la posibilidad de que la selección evolutiva haya ensayado y descartado algunas combinaciones. Esto da idea de la importancia de calcular la probabilidad de que exista una proteína. También para la Ingeniería de Proteínas, sería interesante saber si una combinación de aminoácidos existe en la naturaleza, ó no, con el fin de procurar combinaciones de aminoácidos que la naturaleza ha “aceptado” en proteínas. También consideramos importante identificar las combinaciones de aminoácidos que puedan aparecer en proteínas que presentan condiciones

particulares, como proteínas formadoras de cuerpos amiloides, proteínas con estructuras secundarias muy estables, ó proteínas que permanecen en ambientes extremos, etc.

La estimación de la probabilidad de existencia de una proteína, calculada con los aminoácidos individuales, lleva a obtener probabilidades por encima de cero, mientras que si se calcula la probabilidad de existencia de una proteína usando secuencias cortas, es más factible encontrar secuencias con probabilidad cero de existir en la actualidad, lo que es mucho más cercano a la realidad en el universo proteico conocido.

Queremos resaltar que el número absoluto de apariciones de estas tripletas, de entre 8 y 31 veces, es muy pequeño, considerando que en esta muestra, el número total de conteos de tripletas, fue de 8'784.695. El número de apariciones de las tripletas de menor frecuencia siguió siendo extremadamente bajo, cuando ampliamos el muestreo a la totalidad del PDB (PDBUniverso), con 37'939.740 tripletas. Esto significa que la muy baja (o muy alta) aparición de algunas tripletas en las proteínas, y por tanto, la aparición baja ó alta, de algunas combinaciones de residuos de aminoácidos, debe ser un factor importante a ser tenido en cuenta a la hora de construir ó rediseñar una proteína. La baja aparición de unas tripletas, implica la baja aparición de secuencias de longitudes mayores.

Es decir, si estimamos el número de posibles combinaciones de 4 residuos que se pueden formar con una tripleta cualquiera XYZ, con 20 residuos, estamos hablando de 8000 posibles combinaciones de 4 residuos de aminoácidos. Al haber sólo unas cuántas tripletas, es evidente que habrá muy pocas tétradas de residuos de aminoácidos, lo que confirma que las proteínas no han “ensayado” todas las posibles combinaciones, sino que ha habido una especie de “selección” de unas cuantas combinaciones de residuos de aminoácidos, que han sido “exitosas”, en la construcción de estructuras proteicas funcionales.

A manera de ampliación de la información anterior, es claro que aún hay muchos millones de combinaciones de aminoácidos que no existen en las proteínas, lo que podría estar ocurriendo por dos posibles razones: la primera, que la evolución aún no ha “ensayado” todas las posibles combinaciones para construir proteínas, y/o segunda, que algunas combinaciones de aminoácidos han sido “ensayadas” en la evolución de estas moléculas, y se hayan “descartado”. Por supuesto, aún hay una tercera posibilidad, y es que muchas combinaciones pueden existir en proteínas, pero simplemente hay muchas proteínas que no se han estudiado. Esto lleva a sugerir que las frecuencias del presente estudio se actualicen periódicamente y que se amplíe a secuencias cortas de 5 y 7 residuos, para verificar la aparición de las secuencias que en este momento no se tienen en proteínas, en la medida en que el número de proteínas conocidas siga aumentando sustancialmente.

La estimación de la probabilidad de existencia de una proteína, calculada con los aminoácidos individuales, lleva al cálculo de cualquier probabilidad por encima de cero. Mientras que si se calcula la probabilidad de existencia de una proteína usando secuencias cortas, es más factible

conocer secuencias con probabilidad cero de existir en la actualidad, lo que es mucho más cercano de la realidad.

De aquí que consideramos importante calcular la probabilidad de que exista una secuencia proteica basándose en secuencias cortas, más que en las probabilidades de aminoácidos individuales. Esto sería de interés para la ingeniería de proteínas y eventualmente para la proteómica, y posiblemente también para la biología sintética. Por ello, hemos usado las frecuencias de aparición de aminoácidos individuales y de tripletas, para hacer el cálculo de probabilidad e la existencia de una proteína (Capítulo 5). En nuestras búsquedas de literatura sobre el estimado de la probabilidad de existencia de una secuencia proteica, no hemos encontrado hasta ahora un artículo al respecto. Es posible que este cálculo probabilístico se haya considerado trivial y sin mayor importancia, pero lo consideramos útil para construir una secuencia proteica, y también consideramos que puede ser depurado para grupos específicos de proteínas.

Comparación entre frecuencias esperadas y frecuencias observadas de tripletas

Si se usan las frecuencias de los residuos de aminoácidos individuales como probabilidades, se puede estimar la probabilidad de aparición de una secuencia proteica. También puede calcularse la probabilidad de aparición de las combinaciones de tres aminoácidos, de que habla este trabajo. Esto implica que la probabilidad de aparición de por ejemplo la triplete Phe-Cys-Trp (FCW), cuya probabilidad sería de $pF(0.0389) \times pC(0.0126) \times pW(0.0137) = 0.00000674$, si tomáramos como punto de partida la frecuencia de los residuos en el PDB), debería ser la misma que la probabilidad de aparición de la triplete en otras combinaciones, como Phe-Trp-Cys (FWC), ó Cys-Phe-Trp (CFW), por hablar de sólo dos ejemplos de las seis combinaciones posibles; sin embargo, la frecuencia de aparición de las diferentes combinaciones de estos tres aminoácidos en la triplete, son diferentes, de acuerdo con las frecuencias observadas.

Dos ejemplos más dramáticos, son: el de la triplete WCW, en que la frecuencia esperada es de 0.0002941, mientras que la frecuencia observada es de 0.00000148. Es decir, casi 200 veces menos frecuente de lo esperado! Y una de las tripletas cuya frecuencia observada es mucho mayor que la frecuencia esperada, como es el caso de AAG, cuya frecuencia observada es de 0.00073787, mientras que la frecuencia esperada es de sólo 0.0000022. Es decir, aparece casi 340 veces más que lo esperado! Mientras que las demás combinaciones de los mismos aminoácidos de las dos tripletas de estos ejemplos (CWW ó WWC, y GAA ó GAG) presentan frecuencias observadas muy diferentes. Esto sugiere que al considerar tripletas de aminoácidos, el orden de la combinación de los tres aminoácidos en la triplete, ó en la secuencia en general, es de vital importancia en la construcción de la proteína, como puede verse en la discusión sobre tripletas opuestas.

Por otro lado, si se usan las frecuencias anteriores como probabilidades, se puede estimar la probabilidad de aparición de una secuencia proteica. Así por ejemplo, se podría estimar qué tan

probable es que aparezca la secuencia de la insulina humana, como el producto de las probabilidades de cada uno de los 51 residuos de aminoácidos que la componen. Esto se amplía en el capítulo 5.

Además de lo anterior, se hizo una prueba de Chi cuadrado (χ^2) para comparar las frecuencias observadas, versus las frecuencias esperadas, calculadas como la multiplicación de las probabilidades de los residuos de aminoácidos individuales que conforman la tripleta.

El valor de Chi cuadrado calculado, fue de 1.0. Para un valor α de 0.001 y 7999 grados de libertad, el valor de Chi cuadrado es: 1, lo que nos confirma que las frecuencias son las mismas.

En la figura 3.2.0.1., puede verse que las frecuencias observadas no son iguales a las frecuencias esperadas, calculadas con base en las frecuencias de los residuos de aminoácidos que las componen. Con esto, consideramos que se puede estimar la probabilidad de existencia de una secuencia proteica, con base en las tripletas que la componen. (Ver capítulo 5).

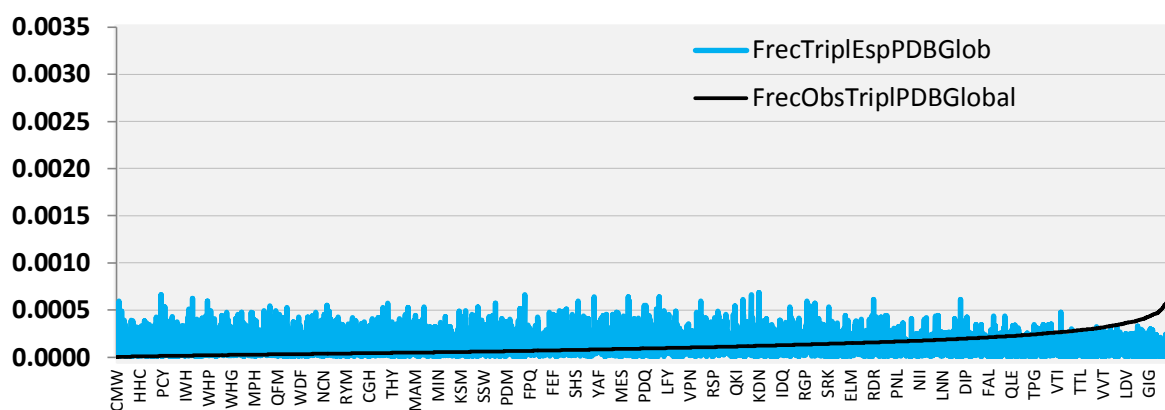


Figura 3.2.0.1. Comparación entre frecuencias de tripletas observadas en el PDBGlobal ordenadas de menor a mayor, y las frecuencias esperadas, para las mismas tripletas, calculadas con las frecuencias de los aminoácidos que las componen, obtenidas de la misma muestra de PDBGlobal.

Como se puede observar en la figura 3.2.0.1., las frecuencias esperadas son más uniformes que las observadas, y además, las observadas son diferentes de las esperadas.

Ya se había reportado que para secuencias de 5 residuos, pentapletas, las frecuencias observadas son diferentes de las frecuencias esperadas (Otaki et al., 2005). Estos autores, también observaron que la distribución de las secuencias cortas, que llamaron “secuencias constitutivas”, no es aleatoria, y que además, muchas secuencias cortas (como las pentapletas), no se encontraron en proteínas. Al sintetizarlas tanto mediante síntesis química de péptidos como por ingeniería genética, verificaron que no hay barreras de tipo energético, ó biológico/genético para la síntesis de estas secuencias cortas en proteínas. Las distribuciones de frecuencias de aparición de las secuencias cortas en proteínas, obedecen a otros factores, tal vez como “unidades para la construcción de proteínas” y lo que las secuencias cortas pueden aportar a la estructura proteica.

En nuestra muestra, muchas tripletas están lejos del valor esperado, pero algunas más que otras. Como ejemplo, la triplete más abundante en la muestra PDBGlobal, es la triplete de tres histidinas (HHH), que es una de las tripletas que más se alejan de su valor esperado (0.003215 observado, vs 0.0000696 como valor esperado; es decir, aparece 46 veces más de lo esperado). Este alejamiento de HHH de su valor esperado, ya había sido observado por Betancourt y Skolnick (2004), en una muestra de 755,973 tripletas en que estudiaron los ángulos diédricos de algunas secuencias en proteínas. No logramos descartar, sin embargo, la posibilidad de que el motivo HHH forme parte de algunas proteínas, como fragmento añadido en una técnica para facilitar el proceso de purificación de la proteína, conocido como “Etiqueta de Histidinas” (“His tag”) (Hochuli et al., 1988; Hengen, 1995; Gavin et al., 2002; Howarth et al, 2006; Zhao et al, 2010).

Las diez tripletas más frecuentes en PDBGlobal, además de la triplete HHH, son: AAA, con frecuencia de 0.001075; AAL:0.000945; ALA:0.000873; EAL:0.000767; AAG:0.000738; ALL:0.000709; AAV:0.000683; AVA:0.000665. Como puede notarse, están compuestas principalmente por los dos aminoácidos más frecuentes en proteínas: Alanina y Leucina, con pocas excepciones.

El hallazgo de secuencias cortas cuya frecuencia es mucho mayor de lo esperado, puede implicar que pequeños motivos como la secuencia HHH, son importantes en la estructura y función de algunas proteínas, con lo que a través de este trabajo, estamos identificando motivos de secuencia que posiblemente sirvan como “ladrillos” de construcción de proteínas con un motivo estructural o funcional particular, que dé características específicas a la estructura proteica.

Una vez que identificamos el motivo HHH, optamos por mirar las frecuencias de tripletas en la totalidad del PDB, que llamamos PDBUniverso, con la idea de identificar otras tripletas con comportamientos alejados de lo esperado.

En la figura 3.2.0.2., salta a la vista que hay dos tripletas en el PDBUniverso, que también se alejan mucho de las frecuencias esperadas, que corresponden a las tripletas WWP, y FQD.

Es llamativo que un motivo de frecuencia frecuente, esté compuesto por residuos de aminoácido con poca frecuencia de aparición de manera individual. El motivo WWP, también conocido como motivo WW, o motivo Rsp5, fue identificado en 1994, por varios grupos (Sudol et al, 1994; André y Springael, 1994; Hofmann y Bucher, 1994), y a partir de allí se ha identificado en varios grupos de proteínas no relacionadas, y tienen que ver con la interacción proteína-proteína y proteína-lípido, gracias al carácter no polar del Trp. Este motivo se puede unir a secuencias ricas en Prolina, que usualmente también contienen Tirosinas. También se ha relacionado este motivo con el sistema ubiquitina-proteasoma (Chondrogianni y Gonos, 2012), y al parecer juega un papel en enfermedades como la hipertensión (Schild et al, 1996) y la distrofia muscular (Huang et al, 2000). Es de resaltar, que de este motivo pueden aparecer múltiples copias en una proteína de

interacción con otras proteínas (Staub y Rotin, 1996). Sobre este motivo, volveremos en el siguiente capítulo, donde presentaremos su asociación con la estructura.

Sobre la otra tripleta que en el PDBGlobal se aleja de la frecuencia esperada, es el motivo FQD (Phe-Gln-Glu – Fenilalanina-Glutamina-Glutámico), no encontramos referencia a ella en la literatura, con lo que consideramos podríamos haber encontrado un motivo de secuencia de importancia en la estructura de las proteínas. Esto se discutirá más ampliamente en el siguiente capítulo, en que asociemos las tripletas de secuencia, con la estructura secundaria.

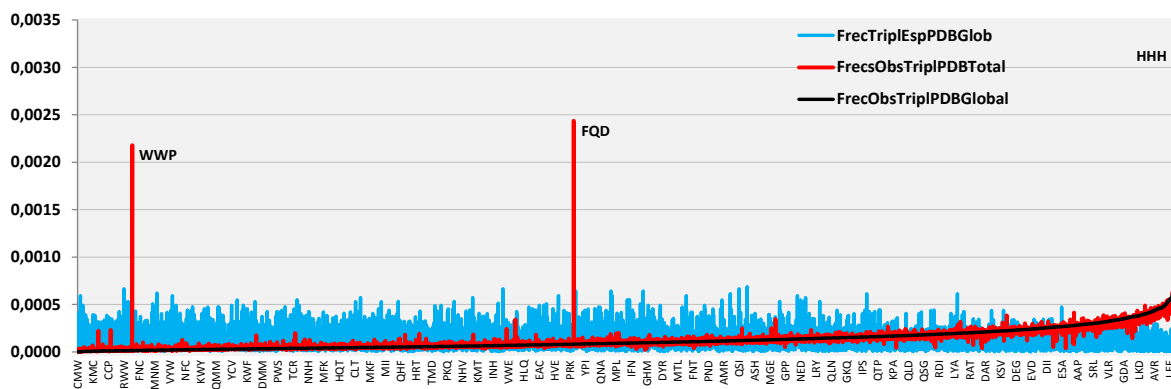


Figura 3.2.0.2. Comparación entre frecuencias observadas en tripletas de residuos de aminoácidos en el PDBGlobal ordenadas de menor a mayor, y las frecuencias observadas en la totalidad del PDB. En color azul, las frecuencias esperadas.

Debido a que la figura anterior sólo evidencia grandes diferencias para tres tripletas, posiblemente debido a la escala que estas tres tripletas fuerzan en la figura, presentamos a continuación la misma figura (figura 3.2.0.3), en que hemos cambiado la escala para mostrar que hay un gran número de otras tripletas que también presentan diferencias en las frecuencias observadas, con respecto a las esperadas.

En la figura 3.2.0.3, se observan diferencias en las frecuencias de otras tripletas, con respecto a la frecuencia esperada. Las más conspicuas se siguen discutiendo en las siguientes páginas.

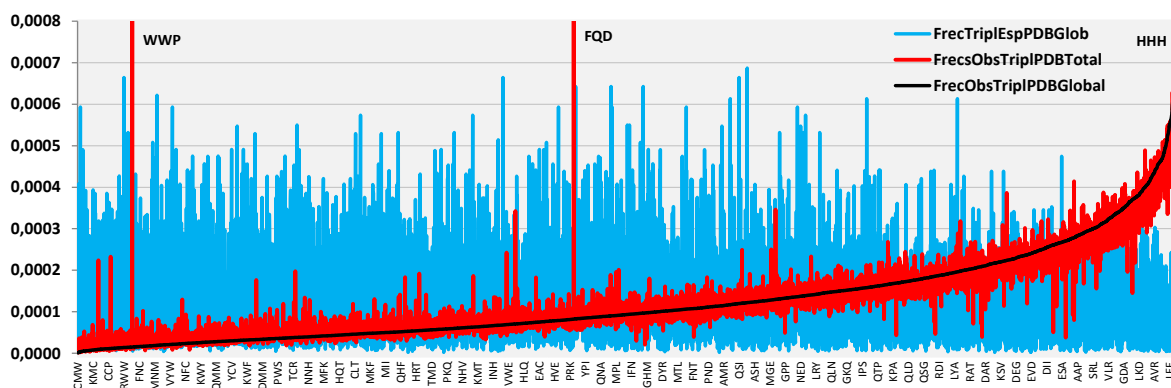


Figura 3.2.0.3. Comparación entre frecuencias observadas en tripletas de residuos de aminoácidos en el PDBGlobal ordenadas de menor a mayor, y las frecuencias observadas en la totalidad del PDB. En color azul, las frecuencias esperadas.

Comparación entre las tripletas de aminoácidos y sus opuestas en secuencia

En esta sección se compararon las frecuencias de las tripletas, con la combinación de residuos de aminoácidos que corresponde al opuesto en secuencia de la tripleta. El fin de este análisis, es el de confirmar, o no, que el orden de los residuos en las combinaciones de aminoácidos, es importante, y se presenta con una frecuencia diferente para las tripletas y sus opuestas en secuencia. En la figura 3.2.0.4. se muestran estos resultados de las frecuencias de las 8000 tripletas, y las de sus opuestas.

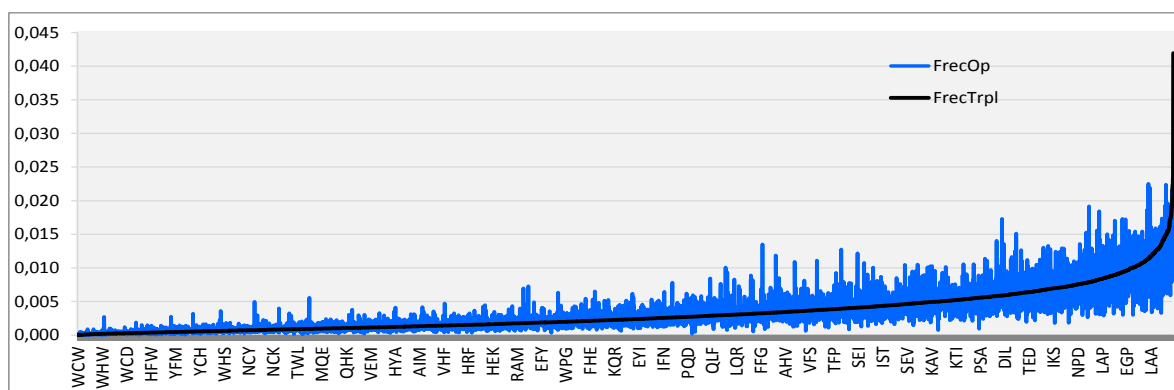


Figura 3.2.0.4. Frecuencias de las 8000 tripletas, ordenada ascendentemente y de sus opuestas en secuencia.

En la figura anterior, se puede observar que no necesariamente las tripletas y sus opuestas, presentan frecuencias similares, lo que confirma el hecho de que la combinación de aminoácidos y el orden de los mismos en las secuencias cortas que constituyen las proteínas, son importantes, y no interreemplazables.

No se puede perder de vista, que de las 8000 tripletas de residuos de aminoácidos, hay 400 tripletas palindrómicas (iguales a su tripleta opuesta), que por tanto no presentan diferencia entre la frecuencia de la tripleta y su opuesta.

Las tripletas y la hidropatía

En la figura 3.2.0.5. se presentan las 8000 tripletas, ordenadas por su hidropatía promedio ponderada, calculada a partir de los residuos que constituyen la tripleta así: Se consideró que el aminoácido central aporta el 50% de su característica a la tripleta, y los dos aminoácidos a lado y lado, aportan cada uno el 25%.

En la figura 3.3.0.5. se puede notar que la hidropatía promedio ponderada es diferente para casi todas las tripletas, y que su valor puede ser importante en términos del aporte que hace a la hidropatía de la proteína. Esto muestra que se tiene una gama de ladrillos de construcción de proteínas, ó como el investigador Joji Otaki las llama, “tripletas constitutivas” (Otaki et al, 2005), con hidropatías variadas, y continuas normalizadas entre -1 y 1.

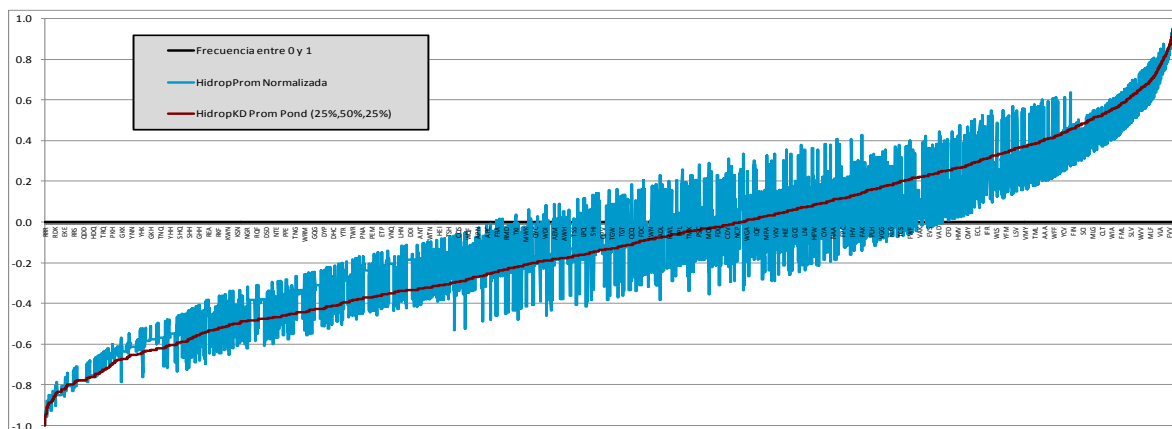


Figura 3.2.0.5. Hidropatías (Kyte & Doolittle, 1982) de las tripletas. En rojo, el promedio ponderado (con respecto al aminoácido central (25%, 50%, 25%), y normalizado. En azul, el promedio simple y normalizado de las hidropatías de las tripletas. En negro, la frecuencia de aparición de cada una de las tripletas.

Las tripletas y sus frecuencias de aparición en proteínas del PDBGlobal

En esta subsección, se han analizado las frecuencias de aparición de las 8000 posibles tripletas de aminoácidos, y tripletas construidas con el alfabeto reducido usado en la sección 2.1. anterior. Usando este alfabeto reducido, se redujo de 8000 a 729 tripletas de letras que representan propiedades particulares de residuos de aminoácidos.

Como primera anotación sobre la aparición de tripletas, a diferencia de los dos trabajos anteriores, que reportaban algunas tripletas que no aparecían en el PDB (Otaki et al, 2003; Lareo y Acevedo, 2004), en el presente trabajo encontramos todas las combinaciones de tres aminoácidos, siendo la tripla más frecuente, HHH, y la menos frecuente, CMW, que apareció sólo 8 veces, de entre más de 8 millones de conteos totales de las tripletas. Las combinaciones menos frecuentes en el PDBGlobal, corresponden principalmente a tripletas de las combinaciones de los residuos menos frecuentes Cisteína (C), Triptófano (W) y Metionina (M).

La tripla más frecuente es HHH construida con un aminoácido que no es de los más frecuentes. Esto puede explicarse en parte porque cerca del 10% de su frecuencia, se debe al uso de la secuencia de seis Histidinas, usado para la expresión y purificación de proteínas en laboratorio, según confirmamos al mirar los archivos PDB de las proteínas de nuestra muestra.

Aparte de ella, las tripletas más frecuentes, que se presentan en la tabla 3.2.0.3, están formadas por los residuos más frecuentes, Alanina (A), Leucina (L), Glicina (G), Valina (V) y Glutámico (E).

En la figura 3.2.0.6., se muestran las frecuencias altas más conspicuas tanto en las tripletas completas, como en las tripletas construidas con el alfabeto reducido.

Tabla 3.2.0.2. Frecuencias de las tripletas menos frecuentes en PDBGlobal.

Total	Tripl	FrecTripAa
8	CMW	0,00000091
11	CCW	0,00000125
12	CHW	0,00000137
13	MWM	0,00000148
13	WCW	0,00000148
14	WCC	0,00000159
14	WWW	0,00000159
15	MWC	0,00000171
16	MWW	0,00000182
18	CIW	0,00000205
19	CMC	0,00000216
21	HWC	0,00000239
23	WCM	0,00000262
25	CCM	0,00000285
25	MWH	0,00000285
27	MCW	0,00000307
28	WWH	0,00000319
31	WCH	0,00000353
31	WPC	0,00000353
31	WWM	0,00000353

Tabla 3.2.0.2. Frecuencias de las tripletas menos frecuentes en PDBGlobal.

Total	Tripl	FrecTripAa
28240	HHH	0,00321468
9443	AAA	0,00107494
8303	AAL	0,00094517
7666	ALA	0,00087265
7401	LAA	0,00084249
6740	EAL	0,00076724
6482	AAG	0,00073787
6225	ALL	0,00070862
6000	AAV	0,00068301
5837	AVA	0,00066445
5829	LLA	0,00066354
5818	LAE	0,00066229
5817	AGA	0,00066217
5795	VAA	0,00065967
5651	ALE	0,00064328
5611	GGG	0,00063872
5547	ALG	0,00063144
5536	EEL	0,00063019
5413	LLE	0,00061619
5406	ELL	0,00061539

Como ya se mencionó, las tripletas más frecuentes, con la excepción de HHH, están construidas con los residuos más frecuentes. Cuando se agrupan en un alfabeto reducido, se hace evidente

que las tripletas más frecuentes en las proteínas del PDBGlobal, son las tripletas construídas con los aminoácidos apolares, y eventualmente con combinaciones de estos y aminoácidos básicos (B: Arginina, Lisina, Histidina), ácidos (X: Aspártico y Glutámico), y aminoácidos con cadena lateral hidroxílica (O: Serina y Treonina).

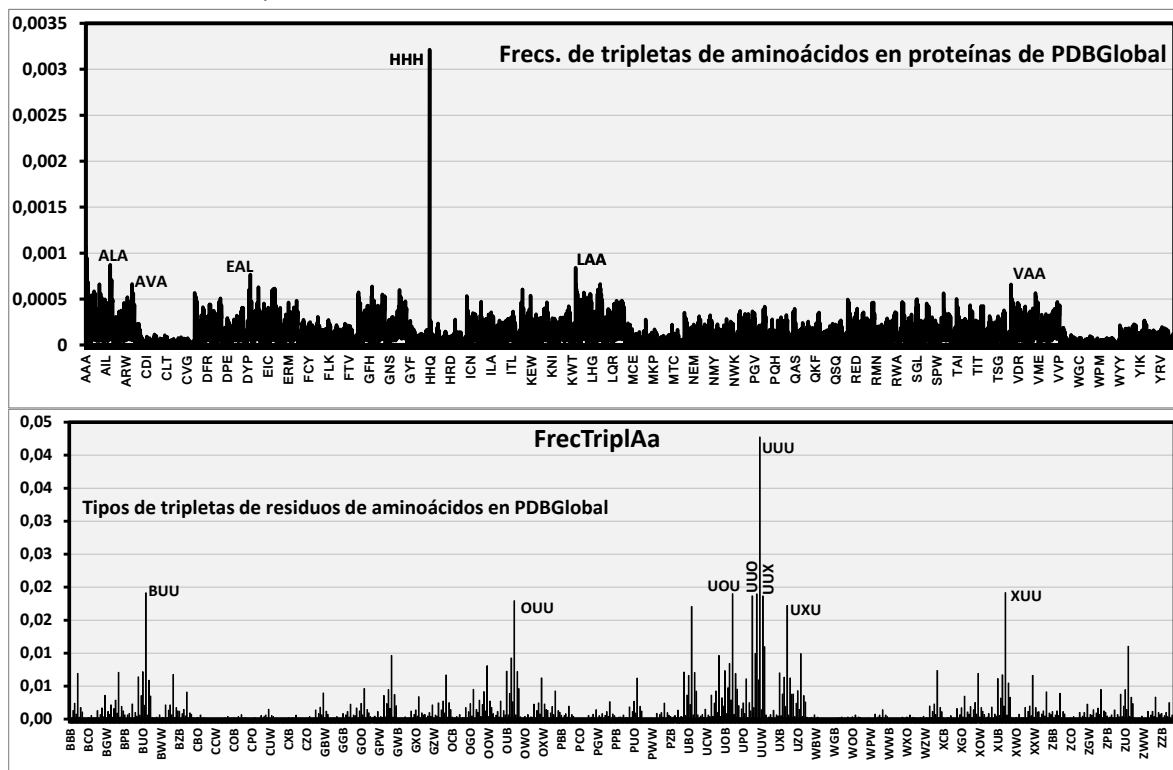


Figura 3.2.0.5. Frecuencia de tripletas por alfabeto reducido en el PDBGlobal. Arriba: tripletas completas. Abajo: tripletas reducidas.

En la figura 3.2.0.5., se puede ver que las tripletas más frecuentes en el PDB Global, son las constituidas, por los aminoácidos hidrofóbicos, UUU, posiblemente por su importancia para la formación del corazón hidrofóbico, que al parecer empuja el plegamiento mediante el colapso hidrofóbico y la formación de un núcleo de residuos apolares al interior, sin importar si la superficie es hidrofóbica ó hidrofílica.

La siguientes tripletas más frecuentes, en orden descendiente, son: la tripleta formada por un residuo ácido, con cadena lateral cargada negativamente, y dos residuos hidrofóbicos, de la forma XUU, su opuesta, de la forma UUX y la forma UXU; la tripleta formada por un residuo básico con cadena lateral cargada positivamente, y dos residuos hidrofóbicos, de la forma BUU, su opuesta UUB y la combinación UBU; la tripleta formada por un residuo de cadena lateral hidroxílica y dos apolares, de la formas UOU, OUU, y UUU.

Las tripletas menos frecuentes son las tripletas formadas por los residuos menos frecuentes de manera individual: C, W y eventualmente, P.

3.2.1. Frecuencias de tripletas en proteínas clasificadas por su composición de estructura secundaria, según SCOP

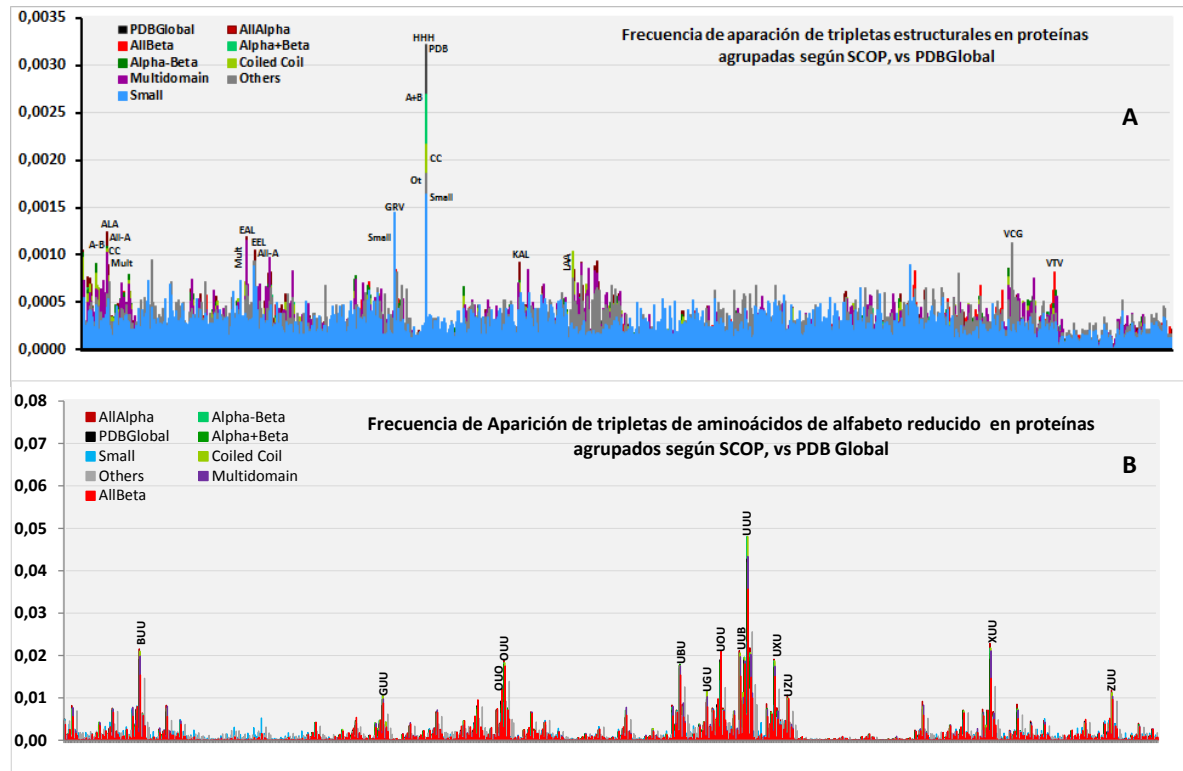


Figura 3.2.1.1. Frecuencias de tripletas de residuos en proteínas agrupadas según SCOP.

En la figura anterior 2.2.1.1., saltan a la vista las tripletas de mayor frecuencia. La tripla HHH que es alta (la más frecuente) en PDBGlobal, también es la más frecuente en proteínas Alfa+Beta, Coiled-Coil, Others, y Small. Ya se comentó en la sección anterior, sobre la mayor frecuencia de esta tripla en muchas proteínas, pero también se dejó abierta la posibilidad de que los investigadores que han reportado estructuras proteicas al PDB, hayan incluido la etiqueta de Histidinas en el extremo amino de las proteínas, con lo que esto podría haber contribuido a la alta frecuencia de esta tripla. Sin embargo, al revisar las secuencias de las proteínas del PDB, sólo cerca del 0.4% de las proteínas, presentan la secuencia de seis Histidinas, que se usa para la purificación de proteínas recombinantes. Esto implica que aunque haya algunas proteínas con las hexa-Histidinas en el extremo amino, estas no son suficientes para explicar la alta frecuencia.

En proteínas pequeñas, la tripla GRV es más alta que las demás tripletas, exceptuando HHH, pero no es la más alta en otros grupos de proteínas. La Glicina en proteínas small, es un residuo muy frecuente, por el contenido de giros que estas proteínas presentan. Los residuos de Arginina y Valina, son residuos frecuentes, aunque no los más frecuentes.

La tripla ALA, es la más alta en proteínas Todo-Alfa, lo que no es de extrañar, ya que los residuos de Alanina y Leucina son fuertes formadores de estructuras helicoidales, según Chou y Fasman, y

las proteínas Todo-Alfa son proteínas ricas en hélices. La misma tripleta ALA, es alta en proteínas “Coiled coil”, y multidominio, que contienen hélices Alfa (lo que se discutirá en el siguiente capítulo).

La tripleta EAL, que es una de las más frecuentes en el PDBGlobal, también es una tripleta frecuente en proteínas Multidominio, y también está formada por residuos de aminoácidos formadores de hélices. Otras tripletas frecuentes como VCG en proteínas “otras”, puede estar hablando de la formación de puentes de disulfuro en proteínas, y VTV en proteínas TodoBeta está mostrando el residuo más fuerte formador de hilos Beta, Valina, y otro fuerte formador de la misma estructura, Treonina. La tripleta IAA es la tercera más frecuente en proteínas tipo Coiled coil.

En la figura 3.2.1.1., se puede ver que el comportamiento en términos de las tripletas más frecuentes, es aproximadamente el mismo que el PDBGlobal. Sin embargo, las proteínas con diferente contenido de estructura secundaria, presentan tripletas diferentes como las más frecuentes: a manera de ejemplo, la tripleta más frecuente para las proteínas Todo-Alfa, son UUU, BUU, UUB y UXU. Para las proteínas Todo-Beta, son UUU, UOU, UUU y OUU. Para las proteínas Alfa+Beta, son las tripletas UUU, BUU, UUB, OUU, y UXU. Para las Coiled-coil, son las tripletas UUU, BUU, UUB. Para las multidominio, UUU, y XUU. Para las proteínas “Otras”, UUU y UOU. Para las proteínas pequeñas, UUU y XUU.

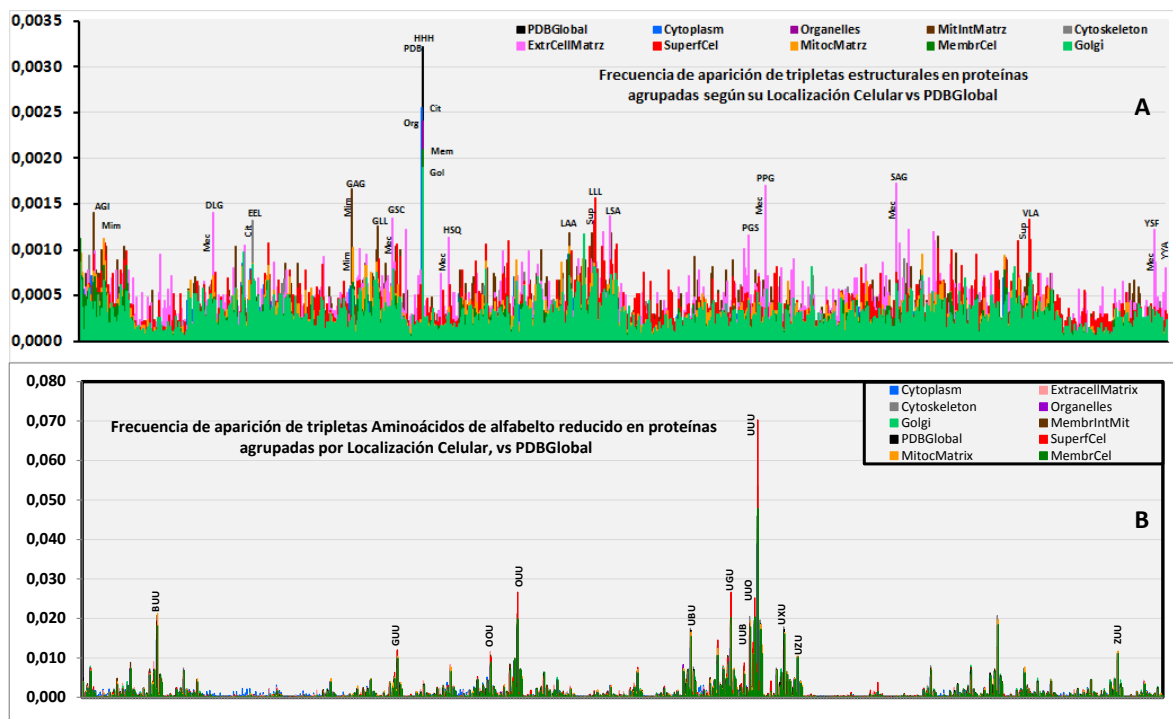
En la figura 3.2.1.1., se puede ver que el comportamiento en términos de las tripletas más frecuentes, es aproximadamente el mismo que el PDBGlobal. Sin embargo, las proteínas con diferente contenido de estructura secundaria, presentan tripletas diferentes como las más frecuentes: a manera de ejemplo, la tripleta más frecuente para las proteínas Todo-Alfa, son UUU, BUU, UUB y UXU. Para las proteínas Todo-Beta, son UUU, UOU, UUU y OUU. Para las proteínas Alfa+Beta, son las tripletas UUU, BUU, UUB, OUU, y UXU. Para las Coiled-coil, son las tripletas UUU, BUU, UUB. Para las multidominio, UUU, y XUU. Para las proteínas “Otras”, UUU y UOU. Para las proteínas pequeñas, UUU y XUU.

Se repite el comportamiento de las proteínas del grupo anterior, en que abundan los residuos apolares, sobre todo en las proteínas de rata y ratón. Lo llamativo es que varias de las tripletas que contienen residuos apolares (“U”), también contienen residuos cargados negativa ó positivamente (de la forma “X” (que agrupa a ácido Aspártico y ácido Glutámico), ó “B” que agrupa a los residuos cargados positivamente Arginina, Histidina y Lisina).

La tripleta formada por tres residuos no polares, resulta la más abundante en proteínas de la superficie celular, aspecto coherente con los tipos de proteínas que interactúan con proteínas de membrana en la célula. Esto ya se ha notado en secciones anteriores, en que los residuos no polares tienden a aparecer en mayor proporción en estas proteínas. Igual ocurre en proteínas del PDB. La frecuencia de aparición de esta tripleta es más de dos veces mayor que las demás tripletas abundantes en estas proteínas de la superficie celular. También es notorio que las tripletas más

3.2.2. Frecuencias de tripletas en proteínas clasificadas por su localización celular.

Las proteínas de la matriz extracelular presentan las mayores diferencias con el PDB, al igual que en el estudio de aminoácidos individuales. Entre ellas, están las tripletas PPG (que forma parte del colágeno, la proteína más abundante de la matriz extracelular). También están en frecuencias altas, las tripletas DLG, CSG, SAG, y PGS, que presentan el residuo de Glicina, asociado con giros y gran libertad de movimiento en los ángulos de giro Psi y Phi, y también asociada con proteínas tipo colágeno.



En las proteínas de superficie celular, la tripleta más abundante es la LLL, formada por un residuo fuerte formador de hilos Beta y también de hélices.

La tripleta más común en proteínas de la membrana interna mitocondrial, es GAG. Esta tripleta, es la de mayor libertad de ángulos de giro Phi y Psi, después de la tripleta GGG, lo que sugiere que podría aportar un motivo de alta flexibilidad a estas proteínas.

Las proteínas del complejo de Golgi, la membrana Celular, y las de organelas, no presentan frecuencias llamativamente diferentes de las frecuencias de tripletas del PDBGlobal.

La tripleta formada por tres residuos no polares, resulta la más abundante en proteínas de la superficie celular, aspecto coherente con los tipos de proteínas que interactúan con proteínas de membrana en la célula. Esto ya se ha notado en secciones anteriores, en que los residuos no polares tienden a aparecer en mayor proporción en estas proteínas. Igual ocurre en proteínas del PDB. La frecuencia de aparición de esta tripleta es más de dos veces mayor que las demás tripletas abundantes en estas proteínas de la superficie celular.

También es notorio que las tripletas más abundantes en las proteínas agrupadas por su localización celular, contengan residuos no polares (agrupados en "U").

3.2.3. Frecuencias de tripletas en proteínas clasificadas por Organismo

En la figura 3.3.3.1., aparece de nuevo la tripleta HHH, como la más frecuente en las proteínas de *Thermotoga*, *Pseudomonas*, *Mycobacterium*, *Bacillus subtilis*, y Humano, incluso por encima de la frecuencia de la misma tripleta en PDBGlobal. La tripleta EAL es la más frecuentemente encontrada en proteínas de *Thermus*, siguiéndole las tripletas ALA, LEA, y EEA, todas ellas con los residuos de Alanina, el más frecuente en nuestra muestra, dos tripletas con Leucina, el segundo aminoácido más frecuente en nuestra muestra del PDBGlobal. Las demás tripletas en los demás grupos de organismos, no presentan diferencias visualmente diferentes que las frecuencias del PDBGlobal.

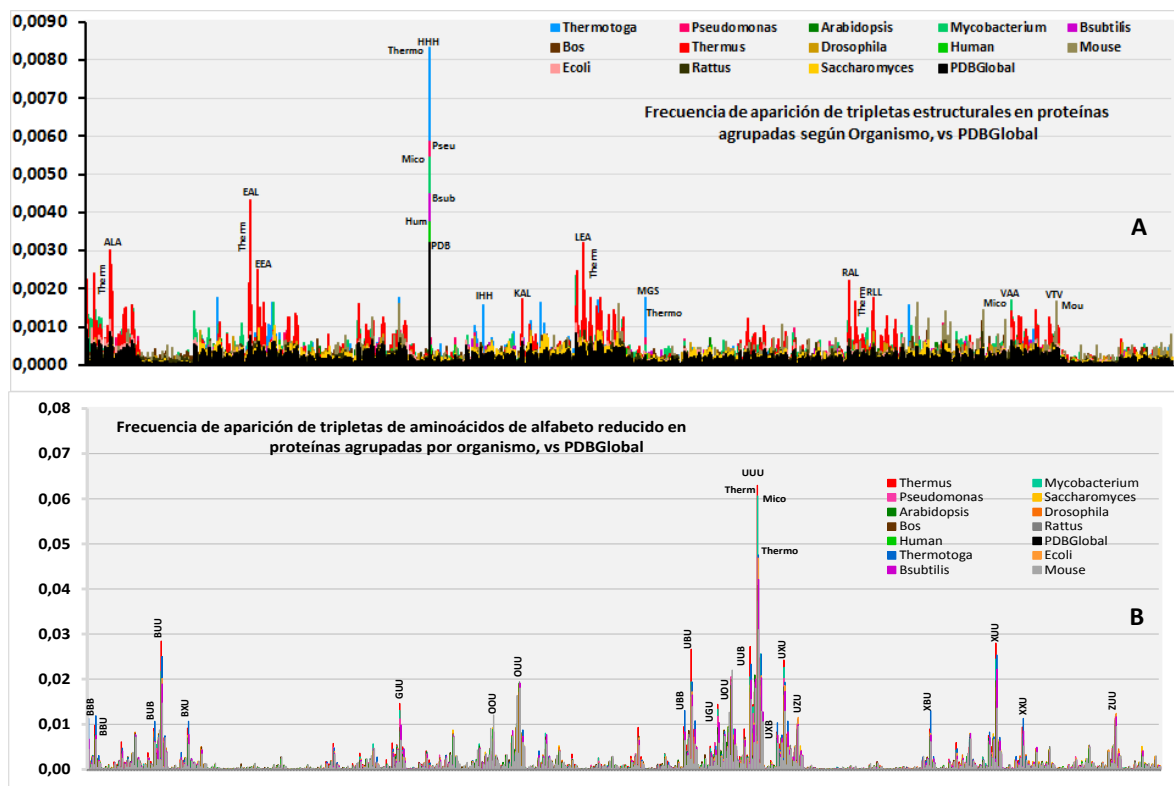


Figura 3.2.3.1. Frecuencias de tripletas de residuos en proteínas agrupadas por Organismo.

Se repite el comportamiento de las proteínas del grupo anterior, en que abundan los residuos apolares, sobre todo en las proteínas de rata y ratón. Lo llamativo es que varias de las tripletas que contienen residuos apolares (“U”), también contienen residuos cargados negativa ó positivamente (de la forma “X” (que agrupa a ácido Aspártico y ácido Glutámico), ó “B” que agrupa a los residuos cargados positivamente Arginina, Histidina y Lisina).

3.2.4. Frecuencias de tripletas en proteínas clasificadas por Phylum

Este fue el grupo de proteínas más homogéneo. De nuevo, aparece la tripleta HHH en mayor frecuencia en las proteínas de Procariota, Archaea, Eucariota y Virus, con frecuencias más altas que las del PDBGlobal en Archaea y Procariotes, y menores que el PDBGlobal, en Eucariota y Virus, como se ve en la figura 3.2.4.1.

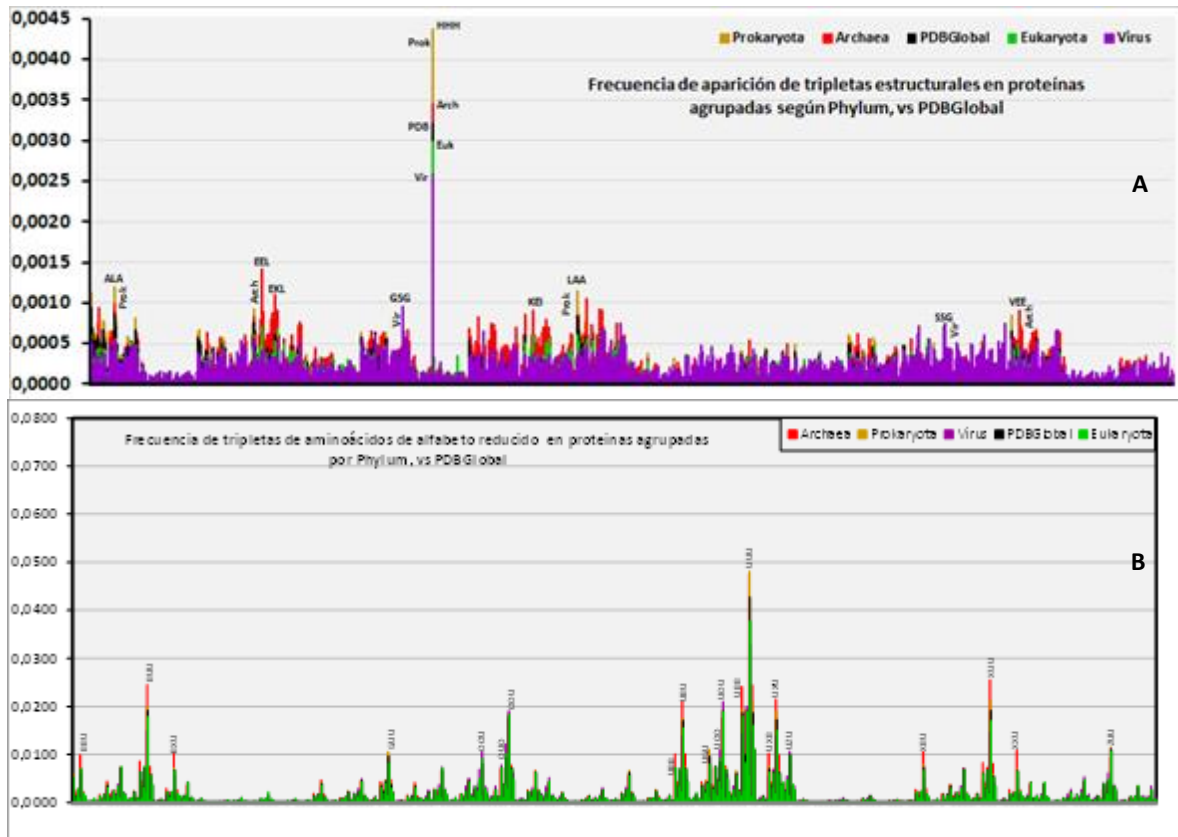


Figura 3.2.4.1. Frecuencias de tripletas de residuos en proteínas agrupadas por Phylum.

La tripleta más común en Archaea es la tripleta EEL; es decir, dos Glutámicos, seguidos por una Leucina. Las más frecuentes en procariota son ALA y LAA, combinaciones de los dos residuos más comunes en proteínas, y formadores preferencialmente de hélices, aunque Leucina, presenta también un alto valor de propensión para hojas Beta. Se observa un comportamiento Igual al de los grupos anteriores.

3.3.5. Frecuencias de tripletas en proteínas clasificadas por Tipo de Enzima

En la figura 3.2.5.1., sigue ocurriendo el mismo comportamiento en la tripleta HHH, frecuente en todos los tipos de enzimas. Con las excepciones de ALA y EAL en Liasas, y LAA e IAA en Isomerasas, son pocas las tripletas que sobresalen en sus frecuencias con respecto al PDBGlobal. Se observa un comportamiento Igual al de los grupos anteriores.

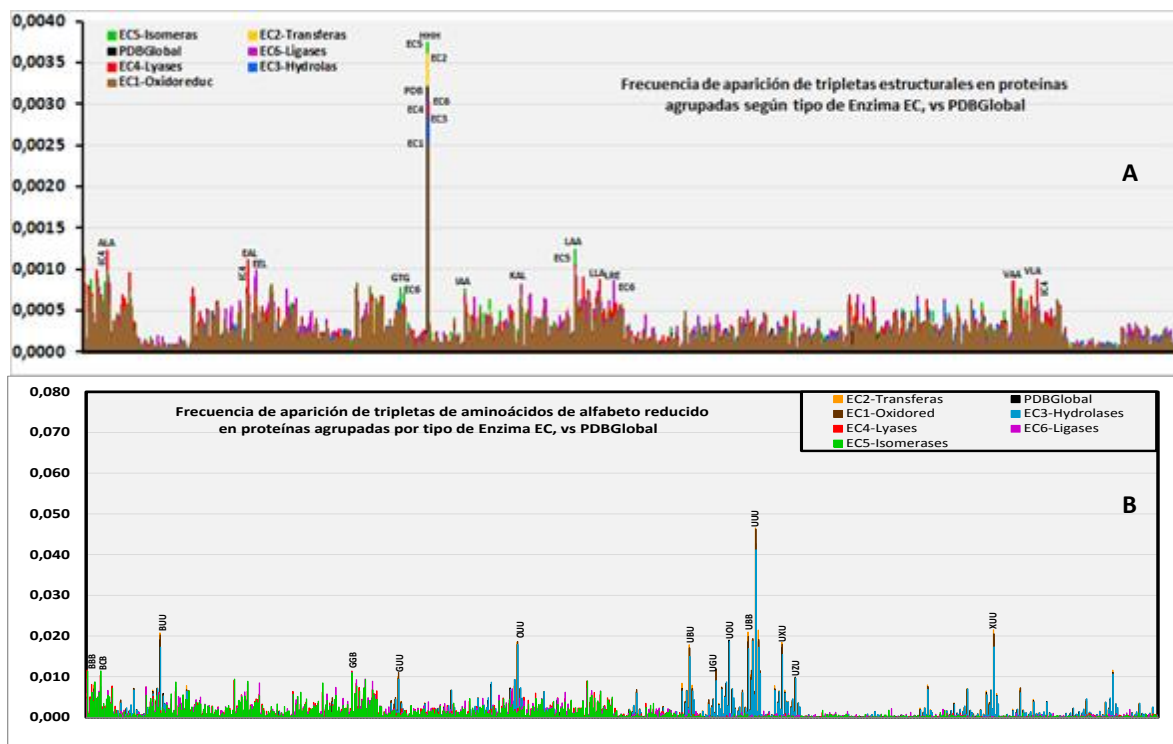


Figura 3.2.5.1. Frecuencias de tripletas de residuos en proteínas agrupadas por tipo de Enzima.

3.2.6. Frecuencias de tripletas en proteínas clasificadas por familia proteica

En la figura 3.3.6.1., es evidente que esta agrupación de proteínas es la que presenta frecuencias más variadas con respecto a la muestra PDBGlobal. La tripleta HHH es común en las proteínas anticongelantes (AFP), las proteínas verde fluorescente y en menor medida, en PDBGlobal. Sin embargo, sí saltan a la vista, tripletas con alta frecuencia de aparición por familias de proteínas en particular: En las proteínas CRISP (Cysteine Rich Secretory Proteins), las tripletas más comunes, son CPS, CPA y YCP, conteniendo Cisteínas y Prolinas. El motivo Cisteína-Prolina es conocido en proteínas con puentes de disulfuro, como las proteínas regulatorias por grupo hemo y en microcolágenos como los encontrados en algunos moluscos, porque la Prolina puede cumplir una función parcialmente oxidante, facilitando el ambiente oxidativo que necesita la Cisteína para el establecimiento de enlaces de disulfuro (Park, 2002; Boulègue, 2006).

En la familia de las Fosfolipasas A2 (PLA2), hay 5 tripletas con frecuencias notoriamente mayores que las demás tripletas, y que las tripletas del PDBGlobal. Son ellas: DRC, RCC, YCG, AIC, CCY, y YCG; todas conteniendo Cisteínas. Las enzimas Fosfolipasas A2 contienen hasta 14 puentes de disulfuro, por lo que es de esperar que contenga tripletas ricas en Cisteínas.

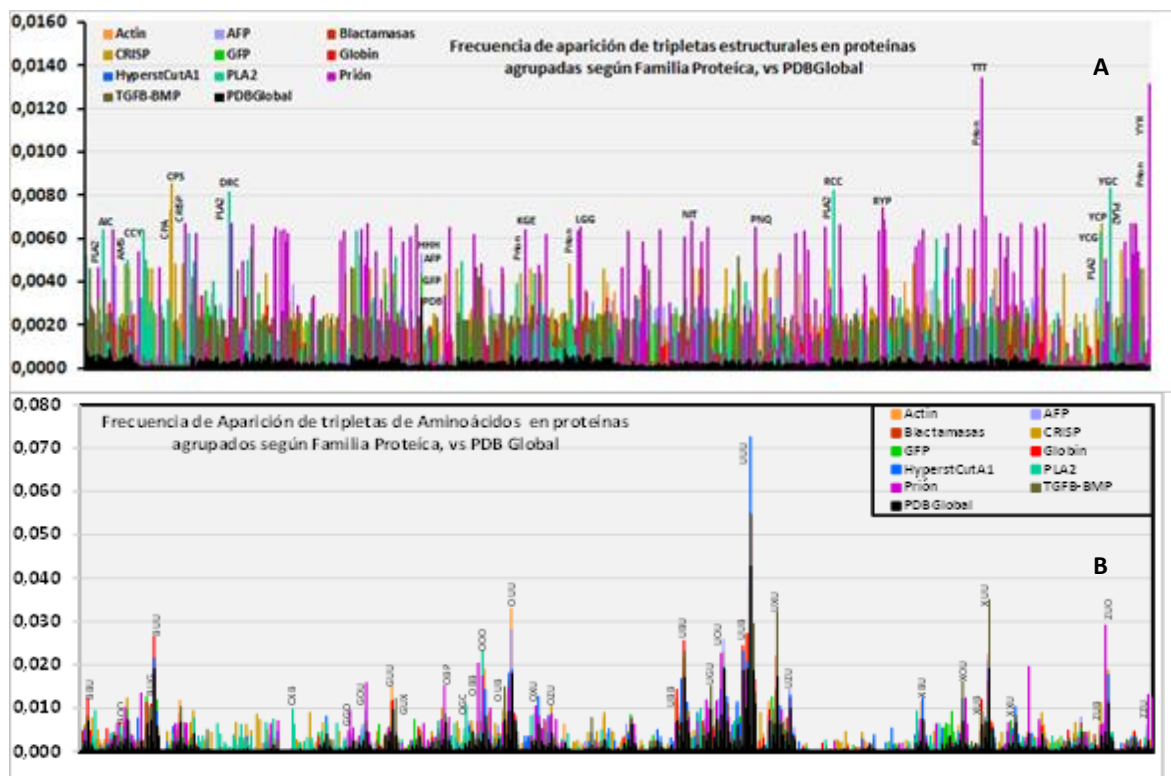


Figura 3.2.6.1. Frecuencias de triplas de residuos en proteínas agrupadas por familia proteica.

La proteína que más salta a la vista por la alta frecuencia de un gran número de triplas, es la familia de priones, en que las triplas con frecuencias llamativamente más altas que las demás, son TTT, y YYR. Estas triplas se discutirán en el siguiente capítulo.

CONCLUSIONES CAP. 3 - FRECUENCIAS DE AMINOÁCIDOS Y TRIPLETAS

Las frecuencias observadas de tripletas en el PDBGlobal, son estadísticamente diferentes de las frecuencias esperadas, calculadas con base en el producto de las frecuencias de aminoácidos individuales. Esto indica que hay independencia entre las frecuencias de los valores individuales y los de las tripletas, y muestra la importancia de estudiar las proteínas en términos de las combinaciones de aminoácidos que componen las secuencias, en lugar de considerar los aminoácidos de forma individual, dando justificación al enfoque utilizado en este proyecto.

Las tripletas de aminoácidos ofrecen un panorama más preciso que lo que se obtiene con los residuos individuales, en términos de la influencia de una secuencia en la estructura de la proteína, y la frecuencia de aparición de tripletas versus la de aminoácidos individuales. La importancia de esto radica en que no son únicamente los residuos individuales los importantes para la construcción de las proteínas, sino las combinaciones de estos residuos, las que determinan las propiedades estructurales de estas moléculas.

La tripleta más frecuentemente encontrada en la muestra PDBGlobal, fue HHH (Histidina-Histidina-Histidina). Aunque se usa una secuencia de seis Histidinas para purificar proteínas clonadas, el número de proteínas con una secuencia añadida de Histidinas, en el PDB, resultó menor del 1%, con lo que restando ese 1% del total de proteínas con este motivo de secuencia, la tripleta HHH aún se puede considerar especialmente frecuente en su aparición natural en proteínas. Sobre esto, se ampliará en un estudio posterior, con la mayor muestra de proteínas disponible, que es la base de datos Uniprot/SwissProt.

Las Cisteínas presentan los valores más bajos en nuestra muestra de la totalidad de las proteínas. Sin embargo, el contenido en este residuo de aminoácido, varía ampliamente entre los grupos de proteínas, sugiriendo que la estabilización de la estructura 3D, y la unión entre subunidades de diferentes proteínas, no siempre se da por enlaces de disulfuro, sobre todo, dado que la principal función de la Cisteína en proteínas, es la de estabilizar la estructura mediante la formación de estos puentes.

La Cisteína aparece con frecuencias notoriamente más elevadas en proteínas de organismos eucariotes, mientras que se presenta en cantidades marcadamente menores, en organismos procariotes, en comparación con la generalidad de las proteínas, lo que indica que los primeros organismos estabilizan sus proteínas soportándose más en puentes de disulfuro, y que los procariotes estabilizan la estructura 3D de sus proteínas, basándose en interacciones débiles, lo cual es esperable, dado el ambiente más reductor del citoplasma de procariotes en comparación con el de eucariotes.

En atención a lo anterior, se puede considerar que el contenido de Cisteína en las proteínas de un organismo, puede usarse como un indicador de la antigüedad evolutiva de ese organismo, lo cual resulta importante para estimar la edad evolutiva de una secuencia de aminoácidos en las proteínas, y por tanto la aparición de esa molécula en el universo proteico

Las Cisteínas, son los residuos de aminoácidos que más varían entre las proteínas agrupadas según los diferentes criterios del presente trabajo, lo cual resulta interesante, dado que la Cisteína es un aminoácido cuyas principales funciones en proteínas, son estabilización de la estructura mediante la formación de puentes de disulfuro, y su participación en mecanismos catalíticos en el sitio activo de algunas enzimas. Por tanto, la variación de Cisteínas habla de la forma como se estabiliza la estructura tridimensional de la proteínas, y eventualmente, de la capacidad catalítica de una enzima

Las frecuencias de aminoácidos y de tripletas en las proteínas agrupadas por tipo de enzima, presentan poca variabilidad, lo que sugiere que la clasificación estructural de enzimas, no se puede realizar con base en la actividad catalítica y tipos de reacciones que realizan.

Las frecuencias de aminoácidos y tripletas son marcadamente diferentes entre familias de proteínas, y diferentes comparadas con la muestra del universo proteico, lo que indica que la clasificación estructural de proteínas es más adecuado si se consideran las proteínas agrupadas por familia proteica.

La agrupación en un alfabeto reducido, evidencia que los aminoácidos hidrofóbicos son los más abundantes en las proteínas, posiblemente por su papel en la iniciación del plegamiento mediante la formación del colapso hidrofóbico.

Las proteínas de pequeño tamaño, tienden a estabilizar su estructura mediante interacciones débiles, dado que muestran contenidos de Cisteínas, especialmente bajos.

El motivo de secuencia WWP se encuentra en una frecuencia anormalmente alta en la totalidad de las proteínas del PDB (PDBUniverso). Este motivo de secuencia, se constituye en un motivo estructural conservado en las proteínas. Por los aminoácidos que componen la tripleta, asume geométricamente la forma de un giro estructural, que resulta en una especie de módulo para la construcción de proteínas. Este giro estructural tiene una geometría particular, dada por el tamaño de la cadena lateral de los Triptófanos, y la torsión que causa la Prolina, por su grupo imino, y se encuentra con frecuencia en proteínas que interactúan con otras proteínas, y tiene aparentemente un papel importante en algunas enfermedades.

La tripleta FQD (Fenilalanina-Glutamina-Glutámico), aparece con una frecuencia llamativamente mayor que la esperada. Sobre esta tripleta, no encontramos referencia en la literatura, con lo que que lo consideramos un motivo de secuencia de importancia en proteínas. Consideramos que este motivo merece un estudio posterior en términos de las estructuras que adopta, para analizar si

ocurre algo parecido al motivo anterior WWP. Su frecuencia especialmente alta, debe estar hablando su importancia como módulo de construcción de algunas estructuras proteicas.

BIBLIOGRAFÍA CAPÍTULO 3 – FRECUENCIAS DE AMINOÁCIDOS Y TRIPLETAS

André B., Springael, J.Y. (1994). **WWP, a new amino acid motif present in single or multiple copies in various proteins including dystrophin and the SH3-binding Yes-associated protein YAP65.** Biochem Biophys Res Commun. 15;205(2):1201-1205.

Axe, D., Foster, N. y Fersht, A. 1996. **Active barnase variants with complexity random hydrophobic cores.** Proc. Nat. Acad. Sci. 93:5590-5594.

Berg OG, Gelb MH, Tsai MD, Jain MK (2001). **Interfacial enzymology: the secreted phospholipase A(2)-paradigm.** Chem. Rev. 101 (9): 2613–54.

Berkmen, M. (2012). **Production of disulfide-bonded proteins in Escherichia coli.** Protein Expression and Purification 82:240–251.

Betancourt, M.R., Skolnick, J. **Local Propensities and statistical potentials of backbone dihedral angles in proteins.** J. Mol. Biol. 342:635-649.

Boulègue, C., Milbradt, A.G., Renner, C., Moroder, L. (2006). **Single proline residues can dictate the oxidative folding pathways of cysteine-rich peptides.** J Mol Biol. 358:846-856.

Brahmachari, S.K., Rapaka, R.S., Bhatnagar, R.S. y Ananthanarayanan, V.S. (1982). **Proline-containing β -turns in peptides and proteins. II. Physicochemical studies on tripeptides with the Pro-Gly sequence.** Biopolymers. 21:1107–1125

Brooks, D.J., Frescos, J.R. (2002). **Increased Frequency of Cysteine, Tyrosine, and Phenylalanine Residues Since the Last Universal Ancestor.** Molec & Cell Proteomics. 1.2:125-131.

Burke, J.E., Dennis, E.A. (2009). **Phospholipase A2 structure-function, mechanism, and signaling.** Journal of Lipid Research. 50: S237–S242.

Calvete, J.J. 2007. **Determinación del número de grupos sulfidrilo y de enlaces disulfuro mediante espectrometría de masas.** Proteómica. 0:21-28.

Cheek, S., Krishna, S., Grishin, N. (2006). **Structural Classification of Small, Disulfide-rich Protein Domains.** J. Mol. Biol. 359, 215–237

Cochran, A.G., Skelton, N.J., Starovasnik, M.A. (2001). **Tryptophan zippers: Stable, monomeric β -hairpins.** Proc. Natl. Acad. Sci. USA 98, 5578.

Chondrogianni, N., Gonos, E.S. (2012). **Chapter 2 – Structure and Function of the Ubiquitin–Proteasome System: Modulation of Components.** En Progress in Molecular Biology and Translational Science. 109:41–74.

- Chu, X.Y., Tian, J., Wu, N.F. y Fan, Y.L. (2010). **An intramolecular disulfide bond is required for the thermostability of methyl parathion hydrolase, OPHC2**. Appl Microbiol Biotechnol. 88(1):125-131.
- Doerr, A. (2009). **Snapshots of the cell surface**. Nature Methods 6, 401:1548-7091
- Du, D., Zhu, Y., Huang C., Gai, F. (2004). **Understanding the key factors that control the rate of β -hairpin folding**. Proc. Natl. Acad. Sci. USA 101, 15915.
- Ferguson, M.A., Williams, A.F. (1988). **Cell-Surface Anchoring of Proteins Via Glycosyl-Phosphatidylinositol Structures**. Annual Review of Biochemistry. 57:285-320
- Fraser ME, Joyce MA, Ryan DG, Wolodko WT. 2002. **Two glutamate residues, Glu 208 alpha and Glu 197 beta, are crucial for phosphorylation and dephosphorylation of the active-site histidine residue in succinyl-CoA synthetase**. Biochemistry. 41(2):537-46.
- Fu H, Grimsley GR, Razvi A, Scholtz JM, Pace CN. (2009). **Increasing protein stability by improving beta-turns**. Proteins. 77(3):491-498.
- Gavin, AC; Bösch, M; Krause, R; Grandi, P; Marzioch, M; Bauer, A; Schultz, J; Rick, JM et al. (2002). **Functional organization of the yeast proteome by systematic analysis of protein complexes**. Nature 415 (6868): 141–7.
- Gustavson, K.H. (1955). **The function of hydroxyproline in collagens**. Nature. 175:70-74.
- Gilbert, H.F. (1990). **Molecular and Cellular Aspects of Thiol–Disulfide Exchange**. Advances in Enzymology 63: 69–172.
- Gilbert, H.F. (1995). **Thiol/disulfide exchange equilibria and disulfide bond stability**. Methods in Enzymology 251: 8–28.
- Gomes-Helene, M.H., Aparecido, P.B., Ponce-Soto, L.A., Marangoni, S. (2012). **Biochemical Characterization and Pharmacological Properties of New Basic PLA2 BrTX-I Isolated from Bothrops roedingeri (Roedinger's Lancehead) Mertens, 1942, Snake Venom**. BioMed Research International. 591470: 1-13
- Hall, D.A. y Reed, R. (1957). **Hydroxyproline and Thermal Stability of Collagen**. Nature 180:243
- Hengen, P (1995). **Purification of His-Tag fusion proteins from Escherichia coli**. Trends in Biochemical Sciences 20 (7): 285–6.
- Hochuli, E.; Bannwarth, W.; Döbeli, H.; Gentz, R.; Stüber, D. (1988). **Genetic Approach to Facilitate Purification of Recombinant Proteins with a Novel Metal Chelate Adsorbent**. Nature Biotechnology 6 (11): 1321–1325.
- Hofmann, K. & Bucher, P. (1995). **The rsp5-domain is shared by proteins of diverse functions**. FEBS Lett. 358, 153–157.

Hopkins FG, Cole SW (1901). **A contribution to the chemistry of proteins: Part I. A preliminary study of a hitherto undescribed product of tryptic digestion.** *J. Physiol. (Lond.)* 27(4–5): 418–28

Howarth, Mark; Chinnapen, Daniel J-F; Gerrow, Kimberly; Dorrestein, Pieter C; Grandy, Melanie R; Kelleher, Neil L; El-Husseini, Alaa; Ting, Alice Y (2006). **A monovalent streptavidin with a single femtomolar biotin binding site.** *Nature Methods* 3 (4): 267–73.

Huang X., Poy F., Zhang R., Joachimiak A., Sudol M., Eck MJ. (2000). **Structure of a WW domain containing fragment of dystrophin in complex with beta-dystroglycan.** *Nat Struct Biol.* 7(8):634-8.

Huber, R., Langworthy, T.A., König, H., Thomm, M. Woese, C., Sleytr, U., Stetter, K. (1986). **Thermotoga maritima sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90°C.** *Archives of Microbiology* 144 (4): 324–333.

Jimenez, S. y Rosenbloom, J. (1974). **Decreased thermal stability of collagens containing analogs of proline or lysine.** *Archives of Biochemistry and Biophysics.* 163 (2): 459–465

Kim, H.W. e Ishikawa, K. (2013). **The role of disulfide bond in hyperthermophilic endocellulase.** *Extremophiles.* 17:593–599.

Kyte, J. y Doolittle, R.F. 1982. **A simple method for displaying the hydropathic character of a protien.** *Journal of Molecular Biology* 157:105-132.

Lareo, L.R. y Acevedo, O.E. 2005. **Amino acid propensities revisited.** *OMICS: A Journal of Integrative Biology.* 9(4): 391-399.

Lo Conte L., Brenner S. E., Hubbard T.J.P., Chothia C., Murzin A. (2002). **SCOP database in 2002: refinements accommodate structural genomics.** *Nucl. Acid Res.* 30(1), 264-267.

Mann, M., Jensen, O. (2003). **Proteomic analysis of post-translational modifications.** *Nat Biotechnol* 21:255-261.

Mejía, D.F., Tischer, I. (2007). **Propuesta de un indicador de la tendencia de una secuencia de residuos de aminoácidos a encontrarse en estructuras secundarias particulares.** Pags 124-130 del Libro : “The 2nd. International Seminar on Genomics, Proteomics, Bioinformatics and Systems Biology”, Casa Editorial de la Universidad del Cauca, Popayan. ISBN 978-958-9451-21-2.

Messens, J., y Collet, J.F. (2006). **Pathways of disulfide bond formation in Escherichia coli.** *The International Journal of Biochemistry & Cell Biology* 38: 1050–1062.

Michelitsch, M.D. y Weissman, J.S. (2000). **A census of glutamineyaspargine-rich regions: Implications for their conserved function and the prediction of novel prions.** *Proc. Natl. Acad. Sci.* 97(22): 11910–11915.

Miseta, A. y Csutora, P. 2000. **Relationship Between the Occurrence of Cysteine in Proteins and the Complexity of Organisms.** *Mol Biol Evol* (2000) 17 (8): 1232-1239.

Moss, G.P. 1992. **Enzyme Nomenclature 1992**. Academic Press, San Diego, California, ISBN 0-12-227164-5, 0-12-227165-3. Suplemento1 (1993), Suplemento2 (1994), Suplemento3(1995), Suplemento4 (1997) y Suplemento5 (en Eur.J.Biochem. 1994, 223, 1-5; Eur.J.Biochem. 1995, 232, 1-6; Eur.J.Biochem. 1996, 237, 1-5; Eur.J.Biochem. 1997, 250; 1-6, y Eur.J.Biochem.1999, 264, 610-650; respectivamente) [Copyright IUBMB].

Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of **proteins database for the investigation of sequences and structures**. J. Mol. Biol. 247, 536-540.

Neves, M.T., Jonson, P.H., Petersen, S.B. (1999). **Amino acid neighbours and detailed information analysis of Cysteines in proteins**. Protein engineering. 12:535-548.

Oka, O.B.V. y Bulleid, N.J. (2013). **Forming disulfides in the endoplasmic reticulum**. Biochimica et Biophysica Acta. 1833: 2425–2429.

Oshima, I., Imahori, K. (1974). **Description of *Thermus thermophilus* (Yoshida and Oshima) comb. nov., a nonsporulating thermophilic bacterium from Japanese thermal spa.** *Int J Syst Bacteriol.* **24**: 102–112.

Otaki, J.M., Shunsuke L., Gotoh, T y Yamamoto, H. 2005. **Availability of short amino acid sequences in proteins**. Protein Science, 14:617-625.

Regeimbal, J. Bardwell, J.C. (2002). **Disulfide bond formation in prokaryotes and eukaryotes**. Chapter 7 in Protein Targeting, Transport & Translocation. ISBN 0-12-200731-X. pp.131-150

Sang-Ho PARK, Hyung-Eun KIM, Chi-Man KIM, Hee-Jeong YUN, Eung-Chil CHOI and Bong-Jin LEE1. (2002). **Role of proline, cysteine and a disulphide bridge in the structure and activity of the anti-microbial peptide gaegurin 5**. Biochem. J. 368:171-182.

Schild L, Lu Y, Gautschi I, Schneeberger E, Lifton RP, Rossier BC (1996). **Identification of a PY motif in the epithelial Na channel subunits as a target sequence for mutations causing channel activation found in Liddle syndrome**. EMBO J. 15 (10): 2381–7.

Seo, J., Lee, K. (2004). **Post-translational modifications and their biological functions: proteomic analysis and systematic approaches**. J Biochem Mol Biol 37:35-44.

Sevier, C.S. y Kaiser, C.A. (2002). **Formation and transfer of disulphide bonds in living cells**. Nature Reviews in Molecular and Cell Biology. 3: 836-847.

Shimoni, L., y Glusker, J.P. 1995. **Hydrogen bonding motifs of protein side chains: descriptions of binding of arginine and amide groups**. Protein Sci. 4(1): 65–74.

Snow, C.D., Qiu, L., Du, D., Gai, F., Hagen S.J., Pande, V.S. (2004). **Trp zipper folding kinetics by molecular dynamics and temperature-jump spectroscopy**. Proc. Natl. Acad. Sci. USA 101, 4077.

Song X, Gragen S, Li Y, Ma Y, Liu J, Yang D, Matoney L, Yan B. (2004). **Intramolecular disulfide bonds are required for folding hydrolase B into a catalytically active conformation but not for maintaining it during catalysis**. Biochem Biophys Res Commun. 319(4):1072-80.

Sudol M, Chen HI, Bougeret C, Einbond A, Bork P. (1994). **Characterization of a novel protein-binding module--the WW domain.** FEBS Lett. 1; 369(1):67-71.

Staub, O., Rotin, D, (1996). **WW domains.** Structure. 4 (5): 495–499

Tsai, J., Gerstein, M., y Levitt, M. (1997). **Simulating the minimum core for hydrophobic collapse in globular proteins.** Protein Sci. 1997 December; 6(12): 2606–2616.

Tu, B.P., Weissman, J.S. (2004). **Oxidative protein folding in eukaryotes: mechanisms and consequences.** The Journal of Cell Biology 164 (3): 341–46.

Witt, D. (2008). **Recent developments in disulfide bond formation.** Synthesis 2008 (16): 2491–2509.

Webb EC., (1992). **Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes.** San Diego: Published for the International Union of Biochemistry and Molecular Biology by Academic Press.

Wojciechowski, C.L., y Kantrowitz, E.R., 2003. **Glutamic acid residues as metal ligands in the active site of *Escherichia coli* alkaline phosphatase.** Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics. 1649(1): 68–73.

Yamada, K.M. (1983). **Cell Surface Interactions with Extracellular Materials.** Annual Review of Biochemistry. 52: 761-799.

Zscherp C., Aygün H., Engels J.W., Mäntele W. (2003). **Effect of proline to alanine mutation on the thermal stability of the all-beta-sheet protein tendamistat.** Biochim Biophys Acta. 1651(1-2):139-145.

Zhao, C; Hellman, LM; Zhan, X; Bowman, WS; Whiteheart, SW; Fried, MG (2010). **Hexahistidine-tag-specific optical probes for analyses of proteins and their interactions.** Analytical Biochemistry 399 (2): 237–45.

CAPÍTULO 4. FRECUENCIAS DE AMINOÁCIDOS Y TRIPLETTAS EN ESTRUCTURAS SECUNDARIAS

En este capítulo 4, se analizan las frecuencias de los aminoácidos individuales y las triplettas en las diferentes estructuras secundarias de DSSP-5. Se compara con la muestra representativa del PDB, y se hacen comparaciones entre agrupaciones de pro

4.1. Frecuencias de residuos de aminoácidos individuales en estructuras secundarias.

En la presente sección, se calculó el contenido de las cinco estructuras secundarias de nuestro interés (Hélices, Hilos y hojas Beta, Doblecetes, Giros Beta, y Estructuras al azar – “random coils”), descritas en el capítulo 0, para compararlo con la muestra representativa del PDB (PDBGlobal), seleccionada según los criterios descritos en el capítulo 1. Las estructuras se obtuvieron por la asignación que hizo el algoritmo DSSP-5 descrito en el capítulo 0, y que usa el PDB como su estándar de oro. Las frecuencias de aparición de aminoácidos en las cinco estructuras secundarias de interés en el presente estudio, se calcularon como el número de veces que aparece cada aminoácido en una estructura secundaria, dividido entre el número total de veces que aparece esa estructura secundaria en cada grupo de proteínas.

Se obtuvieron y analizaron las frecuencias de aparición de aminoácidos individuales en la muestra representativa del PDB y en proteínas agrupadas según diferentes criterios, como contenido de estructura secundaria según la clasificación SCOP, localización celular, organismo en el que se encontraron, Phylum al que pertenecen los organismos en que se encontraron las proteínas, tipo de enzima según la Comisión de Enzimas (EC), y familia proteica según PFam (ver capítulo 2). En todos los casos, se analizaron las frecuencias de los veinte aminoácidos estándar hallados en proteínas y codificados en el código genético (con la excepción de Pirrolisina y Selenocisteína, por estar muy poco representados en el PDB), en cada grupo de proteínas seleccionado.

En general, se ha comparado la composición de estructura secundaria de cada grupo de proteína, con la composición de la misma en la muestra PDBGlobal. También se han comparado las frecuencias de aparición de cada residuo en cada estructura secundaria, entre el PDB y los diferentes grupos de proteínas. Con esto, se han identificado los grupos de proteínas con contenidos de estructuras secundarias particulares, y también los aminoácidos que muestran una frecuencia de aparición mayor en una ó en otra estructura secundaria, y los aminoácidos que aparecen poco en estructuras secundarias, posiblemente tendiendo a desestabilizarlas.

Se identificaron los grupos de proteínas en que aparecen aminoácidos con frecuencias claramente diferentes a la muestra representativa, y se discutió sobre la posible explicación biológica para que

esto ocurra, y eventualmente su utilidad y posible impacto para el rediseño de proteínas y la ingeniería de las mismas.

4.1.1. Frecuencias de aminoácidos en las estructuras secundarias, en proteínas de PDGlobal

Se estudiaron las frecuencias de residuos de aminoácidos en la muestra PDGlobal, en 19.308 proteínas, que sumaron 8'862.961 aminoácidos. Los resultados de las frecuencias de aminoácidos en las estructuras secundarias, se presentan en la tabla 4.tabla 4.1.1.1. y la figura 3.1.1.1.

Tabla 4.1.1.1. Frecuencias de los veinte aminoácidos en las estructuras secundarias en el PDGlobal.

Aa	B	H	S	T	U
A	0,06	0,12	0,06	0,07	0,07
C	0,02	0,01	0,01	0,01	0,01
D	0,03	0,05	0,08	0,08	0,07
E	0,05	0,09	0,06	0,06	0,05
F	0,06	0,04	0,03	0,03	0,03
G	0,05	0,04	0,15	0,19	0,09
H	0,02	0,02	0,03	0,03	0,04
I	0,10	0,06	0,03	0,02	0,04
K	0,05	0,06	0,06	0,06	0,05
L	0,10	0,11	0,05	0,05	0,07
M	0,02	0,03	0,01	0,01	0,03
N	0,03	0,03	0,06	0,07	0,05
P	0,02	0,03	0,05	0,07	0,09
Q	0,03	0,05	0,03	0,03	0,03
R	0,04	0,06	0,05	0,04	0,04
S	0,05	0,05	0,08	0,06	0,08
T	0,07	0,04	0,07	0,04	0,06
V	0,13	0,06	0,04	0,03	0,05
W	0,02	0,02	0,01	0,01	0,01
Y	0,05	0,03	0,03	0,03	0,03

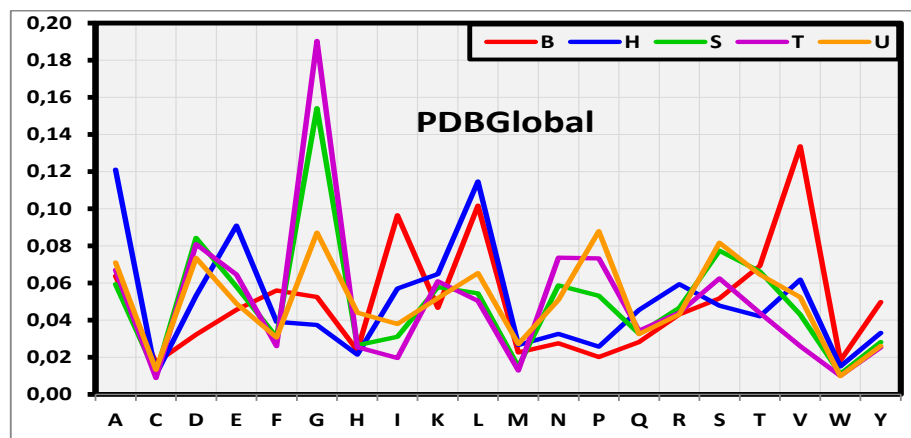


Figura 4.1.1.1. Frecuencia de aparición de aminoácidos en las estructuras secundarias del presente estudio en la muestra PDBGlobal.

Tendencias generales y Diferencias importantes

Los residuos que no aparecen predominantemente en una estructura secundaria:

En PDBGlobal, el residuo de aminoácido menos frecuentemente encontrado en proteínas, la **Cisteína** (C), aparecen con la misma frecuencia, aproximadamente, en cada una de las estructuras secundarias; esto modifica un poco la concepción de que la mayoría de las Cisteínas que no participan en la formación de puentes de disulfuro, tienden a aparecer preferentemente formando parte de hélices (Neves et al, 1999).

Una situación similar ocurre también con el **Triptófano** (W), con lo que estos dos aminoácidos de manera individual, no parecen preferir una estructura secundaria en particular. La **Metionina** (M) y la **Lisina** (K), muestran una variación un poco mayor, pero no una tendencia fuerte hacia ninguna estructura secundaria. Podría pensarse que estos aminoácidos son más dependientes del contexto de secuencia que les rodee para formar una estructura secundaria en particular. Podría decirse que son comodines para la formación de estructuras secundarias.

La **Fenilalanina** (F) muestra una frecuencia ligeramente mayor en hilos Beta, mientras que la **Serina** (S) muestra una ligera tendencia a aparecer en dobles y estructuras al azar. La **Histidina** (H), **Glutamina** (Q), **Asparagina** (N), **Arginina** (R), **Treonina** (T) y **Tirosina** (Y) presentan frecuencias ligeramente más altas en una estructura secundaria, pero sin que esta frecuencia sea notoriamente más alta que en otras estructuras secundarias.

Los residuos que aparecen predominantemente en una estructura secundaria:

Las **estructuras helicoidales** muestran que los aminoácidos que más tienden a formarlas, son la Alanina, el Glutámico y la Leucina. Los hilos/hojas Beta son más fuertemente formados por Isoleucina y Valina. Los giros Beta, son más fuertemente formados por Glicina que por ningún otro aminoácido. Los dobles son formados también por Glicinas, y eventualmente por Serinas. Los

giros al azar son más fuertemente formados por Prolinas y en menor medida por Glicinas y Serinas, pero ningún otro residuo muestra una frecuencia alta en estas estructuras.

La **Alanina** aparece casi el doble de veces en estructuras helicoidales, que en el resto de estructuras secundarias. El ácido **Aspártico** aparece menos en hilos/hojas Beta, que en otras estructuras, sin presentar una frecuencia más alta en ninguna estructura, mientras que el **Glutámico**, aparece preferentemente en hélices, pero sin que sea una frecuencia especialmente alta.

La **Valina** es el residuo más fuerte formador de hilos/hojas Beta, y es tal vez el aminoácido cuya frecuencia de aparición en una estructura, es notoriamente mayor que en las demás estructuras secundarias. Otro aminoácido con frecuencia más alta en esta estructura secundaria, es la Isoleucina.

En la gráfica se puede notar que la **Glicina** (G) aparece más frecuentemente en Giros Beta, y un poco menos en dobleces, y hasta en giros al azar, que en hélices ó en hilos Beta. Según las propensiones de Chou y Fasman y su actualización más reciente, ese aminoácido, tiende a formar giros, lo cual parece ser correcto; sin embargo, las propensiones no identificaron un tipo de giro hacia el cual pudiese presentar una tendencia (Chou y Fasman, 1974a,b,c; Chou y Fasman, 1977, Costantini, 2002). En este análisis, es notorio que su tendencia de las Glicinas, está dada hacia los giros Beta y dobleces.

Por otro lado, la **Prolina**, otro aminoácido con propensión de Chou y Fasman alto hacia la formación de giros, presenta una frecuencia mucho menor que la Glicina, y aparece con mayor frecuencia en giros al azar y en menor medida en giros Beta. También presenta bajas frecuencias en hélices y en hilos Beta.

Aunque la **Leucina** aparece más frecuentemente en estructuras helicoidales, presenta una frecuencia sólo ligeramente más baja en hilos/hojas. Sus frecuencias de aparición en los diferentes tipos de giros, es baja.

4.1.2. Frecuencias de aminoácidos en las Estructuras secundarias, en los grupos de proteínas del presente estudio.

En la presente sección, se analizaron las frecuencias de aparición de los 20 aminoácidos en forma individual en las cinco estructuras secundarias definidas por DSSP (Kabsch y Sander, 1983).

La primera comparación que realizamos, es la de la composición de las cinco estructuras secundarias en las proteínas agrupadas por su contenido de estructura secundaria según SCOP, con la muestra representativa del PDB, que hemos llamado PDBGlobal. Una vez hecho esto, comparamos las frecuencias en que aparece cada aminoácido en cada una de las cinco estructuras secundarias (DSSP-5).

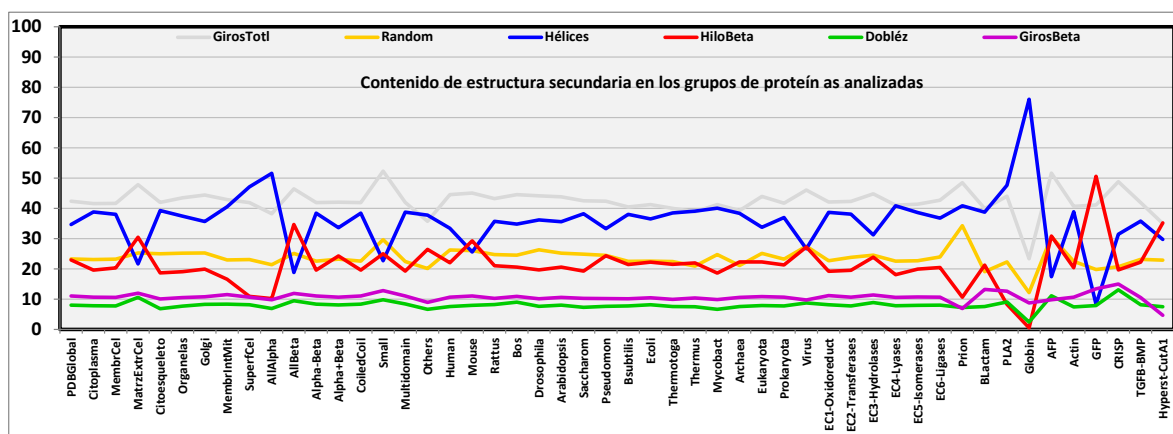


Figura 4.1.2.1. Frecuencia de aparición de las cinco estructuras secundarias en las proteínas del presente estudio (H, B, S, T, U). La suma de los tres tipos de giros, que llamamos “giros totales”, se encuentra en color gris.

En la figura anterior (3.1.2.1.) se puede observar que las proteínas de la muestra PDBGlobal, presentan, por supuesto, los cinco tipos de estructuras secundarias. Sin embargo, hay grupos de proteínas que presentan mayores contenidos de algunas estructuras.

Tendencias generales y Diferencias importantes

La estructura secundaria consistentemente más frecuente, es la estructura helicoidal, que en promedio, aparecen en promedio, en el 36.07% de la composición de estructura secundaria de las proteínas; sin embargo, si se suman los tres tipos de giros en una categoría “giros totales”, esta sería la estructura más frecuente, apareciendo en el 42.43% de la estructura secundaria de las proteínas.

La estructura secundaria que menos aparece en proteínas, es el dobléz, un giro que no tiene un enlace de hidrógeno que contribuya a estabilizarlo y mantenerlo constante, que aparece en el 8.07% de las estructuras secundarias proteicas. La segunda estructura secundaria menos frecuente, que compone el 10.67% de las estructuras secundarias proteicas, es la de los giros Beta, parecidos al dobléz, pero con un enlace de hidrogeno que contribuye a mantener la estructura del giro. Las frecuencias de aparición de dobleces y giros Beta, presentan comportamientos similares en general.

La segunda estructura secundaria más frecuente fluctúa entre los hilos/hojas Beta, y los giros al azar. Aunque los giros al azar constituyen en promedio el 23.69% de la estructura secundaria de las proteínas, en algunas proteínas, predominan las hojas Beta (21.5%), sobre los giros al azar, y en otras predominan estos últimos, sobre los hilos/hojas Beta.

Hay un grupo de proteínas que presenta una cantidad especialmente alta de hélices (76.04%), más que la frecuencia más alta de una estructura secundaria en todas las demás proteínas del estudio.

Este grupo de proteínas es el de las Globinas, que además tienen el menor contenido de hilos Beta (0.57%) y también el menor contenido de dobleces (2.46%), de entre todas las 52 agrupaciones de proteínas de este estudio. Es llamativo que el contenido de hélices de esta proteína, es aún más alto que el de las proteínas que según la clasificación SCOP, corresponde a las proteínas “todo-Alfa”, que es de proteínas predominantemente ricas en estructuras helicoidales. Este grupo de las Todo Alfa presenta el segundo grupo con mayor contenido de hélices, con un 51.60%.

Las proteínas con mayor contenido de hilos y hojas Beta, son las proteínas Verde Fluorescentes (GFP), con un contenido de esta estructura, del 50.57%. También son las proteínas con menor contenido de hélices, con 8.29%. Las GFP son un barril Beta, con dos cortas hélices, una adentro del barril y otra en la parte inferior. Otra proteína con alto contenido de hilos/hojas Beta, son las proteínas Hiperestables CutA1, en que su estructura secundaria contiene un 35.18% de hilos y hojas. Su contenido de hélices es similar, con 29.78%.

Resulta llamativo que las proteínas de los Priones, aunque tienen un contenido de hélices cercano al promedio en las proteínas del estudio, con 40.86%, muestran también el mayor contenido de estructuras de giros al azar, con 34.29%, lo que en parte puede contribuir a explicar la poca estabilidad de su estructura tridimensional y su tendencia a cambiar el plegamiento. Estas proteínas muestran un contenido menor de hilos/hojas Beta que el promedio de las proteínas (10.67%).

Las proteínas clasificadas por SCOP como “pequeñas”, son las únicas que contienen más giros al azar que las demás estructuras secundarias. Es importante recordar, que dentro de este grupo de proteínas, se incluyen algunas de las proteínas clasificadas como “intrínsecamente desordenadas”. Su alto contenido de “random coils” contribuye a explicar su desorden estructural intrínseco.

Aunque en general, las proteínas muestran mayores contenidos de hélices que de hilos/hojas Beta, unas cuantas proteínas invierten esa proporción. Una de ellas, la agrupación de proteínas de la matriz extracelular presentan mayor contenido de hojas que de hélices. Lo mismo ocurre para las proteínas Todo-Beta, y las proteínas anticongelantes (AFP), que contienen igual proporción de hilos/hojas, que de giros al azar, lo que podría explicar parcialmente su estabilidad a muy bajas temperaturas.

Las proteínas agrupadas por organismo, presentan aproximadamente las mismas proporciones de las estructuras secundarias, con la excepción de las proteínas de ratón, que contienen ligeramente mayor contenido de hilos/hojas, que de hélices.

Las proteínas “Coiled coil” presentan frecuencias en sus estructuras secundarias, muy parecidas a las de las proteínas Alfa-Beta (o Alfa/Beta) y hasta cierto punto, las Alfa+Beta, que a su vez se parecen bastante a las frecuencias de la muestra PDBGlobal.

Las proteínas “pequeñas” presentan el contenido más alto de giros al azar (“random coils”) de las proteínas de la figura, y contenidos bajos y casi idénticos de hélices y hojas. Esto muestra que las proteínas “pequeñas” presentan hélices y hojas más cortas, conectadas con giros más largos, con lo que la dominancia relativa en contenido, será de los giros.

Las proteínas Alfa-Beta (ó Alfa/Beta), que están formadas principalmente por hojas Beta paralelas, en unidades (Beta-Alfa-Beta), y las proteínas Alfa+Beta, formadas principalmente por hojas Beta antiparalelas, en regiones segregadas de Alfa y Beta, presentan aproximadamente los mismos contenidos de giros, pero Alfa-Beta presenta mayores contenidos de hélices, mientras que las Alfa+Beta, presentan mayores contenidos de hilos/hojas Beta.

Los giros totales suman en general, entre el 38 y 46% de los contenidos de las proteínas de la figura con las más notorias excepciones en las proteínas llamadas “otras”, que presentan menores contenidos de estas estructuras de giros, en alrededor del 35%, mientras que las proteínas “pequeñas”, presentan los mayores contenidos de estas estructuras, llegando al 53%.

Los dobleces se mantienen entre el 6 y el 10% de la totalidad de estructuras en las proteínas, y los giros Beta fluctúan entre el 9 y el 11%.

La composición de Giros tipo Beta y Dobleces, es aproximadamente la misma en el PDBGlobal y en los grupos de proteínas clasificados según SCOP. Las proteínas presentan contenidos aproximadamente constantes de estas proteínas clasificadas según SCOP, sin importar si contienen más hélices u hojas.

En la figura 3.1.2.1. se observa una distribución bastante homogénea en los contenidos de estructura secundaria en las proteínas agrupadas por organismo, con pequeñas excepciones, como el hecho de que las proteínas de ratón presentan menores contenidos de estructuras helicoidales, y mayores contenidos de hilos/hojas Beta que las proteínas de la generalidad (PDBGlobal). Por el contrario, las proteínas de Mycobacterium presentan los contenidos más altos de hélices y los contenidos más bajos de hilos/hojas Beta.

4.1.3. Frecuencias de aminoácidos en las Estructuras helicoidales, de las proteínas evaluadas en el presente estudio.

En esta sección, se comparan las frecuencias de aminoácidos en las cinco estructuras secundarias del estudio (DSSP-5). En la figura siguiente, 3.1.3.1., se muestran las frecuencias de residuos en las estructuras helicoidales.

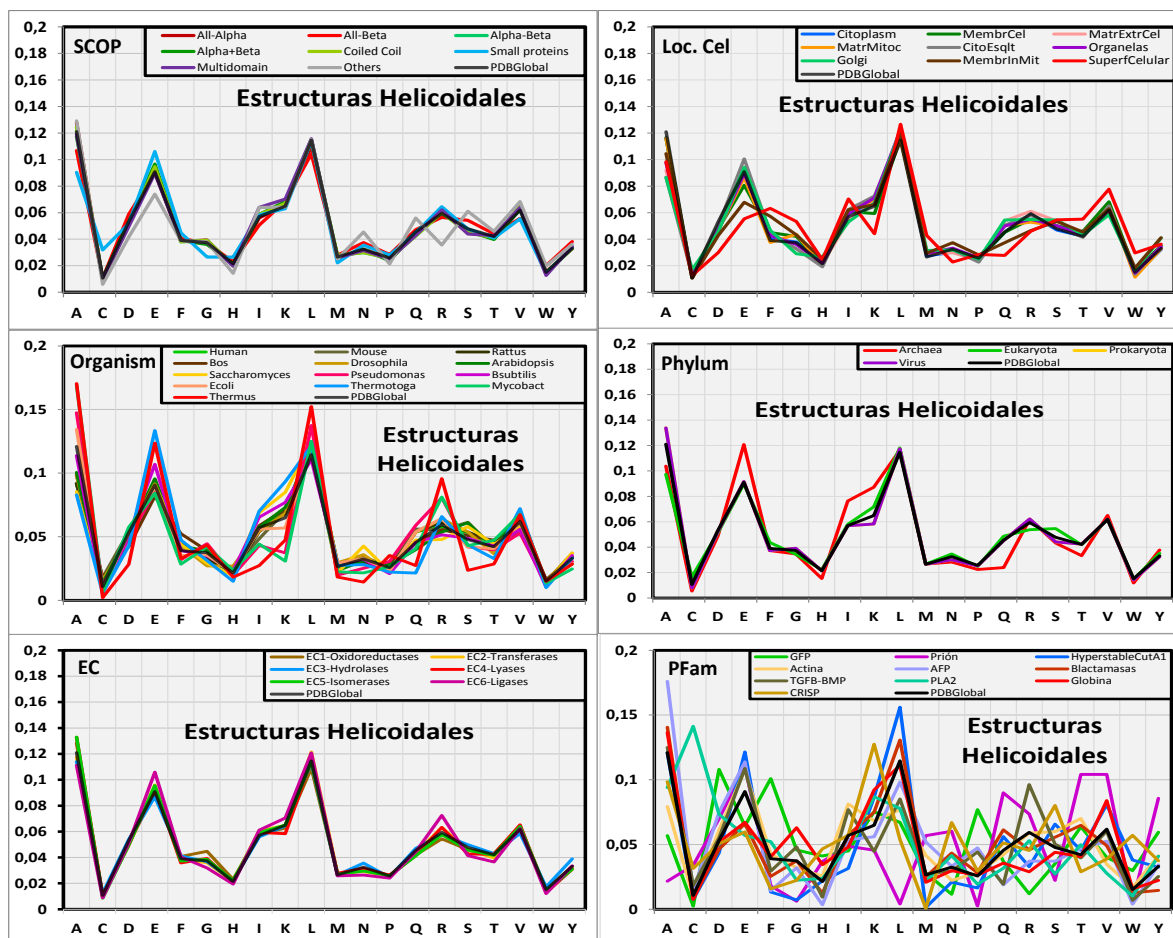


Figura 4.1.3.1. Frecuencia de aparición de aminoácidos en estructuras helicoidales, en las proteínas analizadas en el presente estudio.

Tendencias generales y Diferencias importantes

Las estructuras helicoidales

En la figura anterior se observa que los aminoácidos de las diferentes agrupaciones proteicas, son en general bastante similares a las de la muestra PDBGlobal, con la notoria excepción de las proteínas agrupadas por familia. Es decir, con la excepción de las familias de proteínas, los aminoácidos se comportan aproximadamente igual, en términos de la frecuencia de aparición en estructuras secundarias helicoidales, con algunas excepciones puntuales en las demás agrupaciones proteicas. Las proteínas clasificadas según el tipo de enzima, presentan la mayor homogeneidad en sus estructuras secundarias, lo que sugiere que los plegamientos usados para construir proteínas con actividad catalítica, son aproximadamente los mismos para catalizar diferentes tipos de reacciones. Sugerimos que esto debería confirmarse estudiando las subclases de enzimas, para mirar más finamente si reacciones específicas, presentan una tendencia a un plegamiento en particular.

Los residuos que más frecuentemente forman estructuras helicoidales, son las Alaninas, Leucinas y Glutámicos. Los menores formadores de estas estructuras, son la Cisteína y el Triptófano, seguidos de Histidina, Metionina y Prolina.

Proteínas agrupadas por su contenido de estructura secundaria según SCOP

En las proteínas agrupadas por su contenido de estructura secundaria, las mayores diferencias se pueden encontrar en proteínas “pequeñas”, en los aminoácidos Cisteína, y Glutámico, que aparecen más frecuentemente en hélices que lo que aparecen en las demás proteínas, y el aminoácido Glicina que aparece menos en hélice que las demás. En el capítulo anterior, se observó que la Cisteína se encuentra en una frecuencia especialmente alta en proteínas pequeñas.

Las proteínas clasificadas como “Otras”, también presentan diferencias en los residuos de Glutámico y Arginina, que están menos frecuentemente en hélices, y la Asparragina y Serina, que aparecen de forma ligeramente más frecuente en hélices que lo que aparecen en las demás proteínas. También Arginina y Triptófano se observan más frecuentemente en proteínas clasificadas como “otras”, sobre las frecuencias de los mismos residuos encontrados en las demás proteínas.

Proteínas agrupadas por su localización celular

En este grupo, las proteínas de superficie celular presentan las mayores diferencias con respecto a la muestra PDBGlobal y las demás proteínas de diferentes localizaciones celulares. Las mayores diferencias se pueden considerar así: los residuos de aminoácidos que presentan una frecuencia mayor en hélice que el PDBGlobal, son Fenilalanina, Glicina, Isoleucina, Metionina, Treonina Valina y Triptófano, mientras que los aminoácidos que presentan menor frecuencia son: los aminoácidos ácidos Aspártico, Glutámico, Lisina, Asparragina, Glutamina y Arginina. Mientras que las hélices de estas proteínas de superficie celular son más ricas en Fenilalanina, Glicina, Leucina, Treonina, Valina, y Triptótano. Según esto, las hélices de las proteínas de la superficie celular, tienden a ser hélices menos ricas en residuos polares, y más ricas en residuos no polares; dicho de otra manera, presentan hélices principalmente hidrofóbicas. Esto es coherente con el hecho de que estas proteínas tienden a presentar una fuerte interacción hidrofóbica con proteínas de la membrana y/o proteínas de la matriz extracelular, por lo que de alguna manera disminuyen sus interacciones polares.

También se observan pequeñas diferencias en los aminoácidos Glutámico, Glutamina y Arginina, que aparecen menos frecuentemente en proteínas de la membrana interna mitocondrial, mientras que Fenilalanina y Glicina se encuentran con una frecuencia ligeramente mayor en las mismas proteínas, con respecto a la muestra PDBGlobal.

Proteínas agrupadas por organismo

Las mayores diferencias se presentan en proteínas de las dos bacterias termófilas, *Thermus* y *Thermotoga*, y en menor medida en las proteínas de *Mycobacterium*.

En *Thermus*, la Alanina, Leucina, Glutámico y Arginina, son los residuos con mayores frecuencias en hélices, mientras que Aspártico, Isoleucina, Asparragina, Glutamina, Serina y Treonina aparecen menos en hélices.

En *Thermotoga*, Alanina aparece menos en hélices, al igual que Glutamina. En cambio, el Glutámico y la Isoleucina aparecen más frecuentemente en estructuras helicoidales. Tanto en *Thermotoga* como en *Thermus*, se encuentran particularidades debidas posiblemente a que sus proteínas están expuestas a un ambiente de alta temperatura, lo que las hace interesantes para la búsqueda de secuencias que provean termoestabilidad a las proteínas.

Mycobacterium, muestra una Lisina menos frecuente en estructuras helicoidales. Posiblemente debido al ambiente ácido en que usualmente vive, que podría interactuar fuertemente con la carga positiva de este aminoácido.

Proteínas agrupadas por Phylum

Las mayores diferencias se presentan en proteínas de Archaea, en que Glutámico, Isoleucina, y Lisina aparecen más en hélices que las demás proteínas, mientras que Histidina, Glutamina y Treonina aparecen menos.

Proteínas agrupadas por Tipo de Enzima

Como se ha mencionado antes, el grupo de enzimas es muy homogéneo en cuanto al uso de aminoácidos en estructuras helicoidales. Esto sugiere que los plegamientos usados para construir proteínas con actividad catalítica, son aproximadamente los mismos para catalizar diferentes tipos de reacciones. Sugerimos que esto debería confirmarse estudiando las subclases de enzimas, para mirar más finamente si reacciones específicas, presentan una tendencia a un plegamiento en particular.

Proteínas agrupadas por Familia proteica

En esta agrupación, que presenta la mayor heterogeneidad, las diferencias más notorias están en que las proteínas Hiperestables CutA1 hacen un mayor uso de Glutámico y Leucinas para la construcción de las hélices, y menos uso de Fenilalanina, Glicina, Isoleucina, Metionina, Asparragina, Prolina y Arginina. Dado que Glutámico y Leucina parecen ser fuertes formadores de hélices, según las propensiones de Chou y Fasman, lo que sugiere que sus hélices pueden ser bastante estables.

A manera de contraste, las hélices de los priones, presentan el menor contenido de Leucinas de entre todas las 52 tipos de proteínas del presente estudio, con lo que esencialmente carecen del residuo más estable en hélices, y en cambio están construidas con Glutamina, Arginina, Treonina, Valina y Tirosina. La Valina, según Chou y Fasman, son los residuos mayores formadores de Hilos y hojas Beta (Chou y Fasman, 1974a,b,c;; Chou y Fasman, 1977, Costantini, 2002). Esto puede

contribuir a explicar en parte, la tendencia de las proteínas del Prión, a cambiar la estructura tridimensional, asociada con enfermedad.

En GFP, las hélices aparecen formadas por Glutámico, Fenilalanina, y Prolina, principalmente, mientras que en TGFB-BMP, las hélices se construyen principalmente con Glutámico, Isoleucina, Leucina, y Arginina. Las proteínas CRISP construyen sus estructuras helicoidales en mayor medida, con Lisina, Serina y Triptófano. Las proteínas AFP son las que más construyen sus hélices usando Alanina. También usan Glutámico para ello. Las Fosfolipasas A2 (PLA2), construyen sus hélices usando principalmente Cisteína, algo completamente diferente de la proporción de Cisteínas en cualquier estructura secundaria. La Cisteína presenta una frecuencia de aparición baja en todas las estructuras secundarias, y en el caso de las PLA2, este aminoácido es el que aparece en mayor proporción en estructuras helicoidales.

La construcción de hélices con diferentes aminoácidos en diferentes familias, da cuenta de la necesidad de que la estructura proteica se acomode a las particularidades del microambiente en que están inmersas las proteínas y eventualmente, en la función que realizan.

4.1.4. Frecuencias de aminoácidos en hilos y hojas Beta, de las proteínas evaluadas en el presente estudio

La siguiente figura resume el empleo de los diferentes aminoácidos en **hilos y hojas beta**, en las proteínas agrupadas de acuerdo de los diferentes criterios.

Tendencias generales y Diferencias importantes

Al igual que en las estructuras helicoidales, las proteínas que muestran mayor homogeneidad en la frecuencia de residuos en hilos y hojas Beta, son las proteínas agrupadas por tipo de enzima, y las proteínas más variadas en sus frecuencias en estructuras secundarias son las proteínas agrupadas por familia. Las segundas más variadas, son las proteínas agrupadas por organismo.

Los residuos que más frecuentemente aparecen en estructuras hilos/hojas Beta, son las Valinas, seguidas por las Leucinas e Isoleucinas, todos aminoácidos apolares. Los aminoácidos Cisteína y Triptófano, son los que menos aparecen en estas estructuras secundarias, seguidos por Histidina, Metionina, y Prolina.

Proteínas agrupadas por su contenido de estructura secundaria según SCOP

En la figura anterior (3.1.5.1.), aunque los aminoácidos tienden a comportarse de manera similar en la mayoría de los grupos proteicos comparados con el PDB, es evidente que hay pequeñas excepciones, principalmente provenientes de las proteínas “pequeñas”, y de las proteínas clasificadas como “otras”. En las “pequeñas”, la Cisteína tiende a estar en mayor proporción en estructuras Beta que cuando está en el resto de proteínas de otros contenidos de estructura

secundaria (en el capítulo anterior ya se discutió la importancia de las Cisteínas en la estabilización de proteínas pequeñas). En estas proteínas, la Valina, conocido como el más fuerte formador de hilos y hojas Beta, presenta en este grupo, la menor frecuencia hacia la formación de estas estructuras, con respecto al mismo residuo en las demás proteínas.

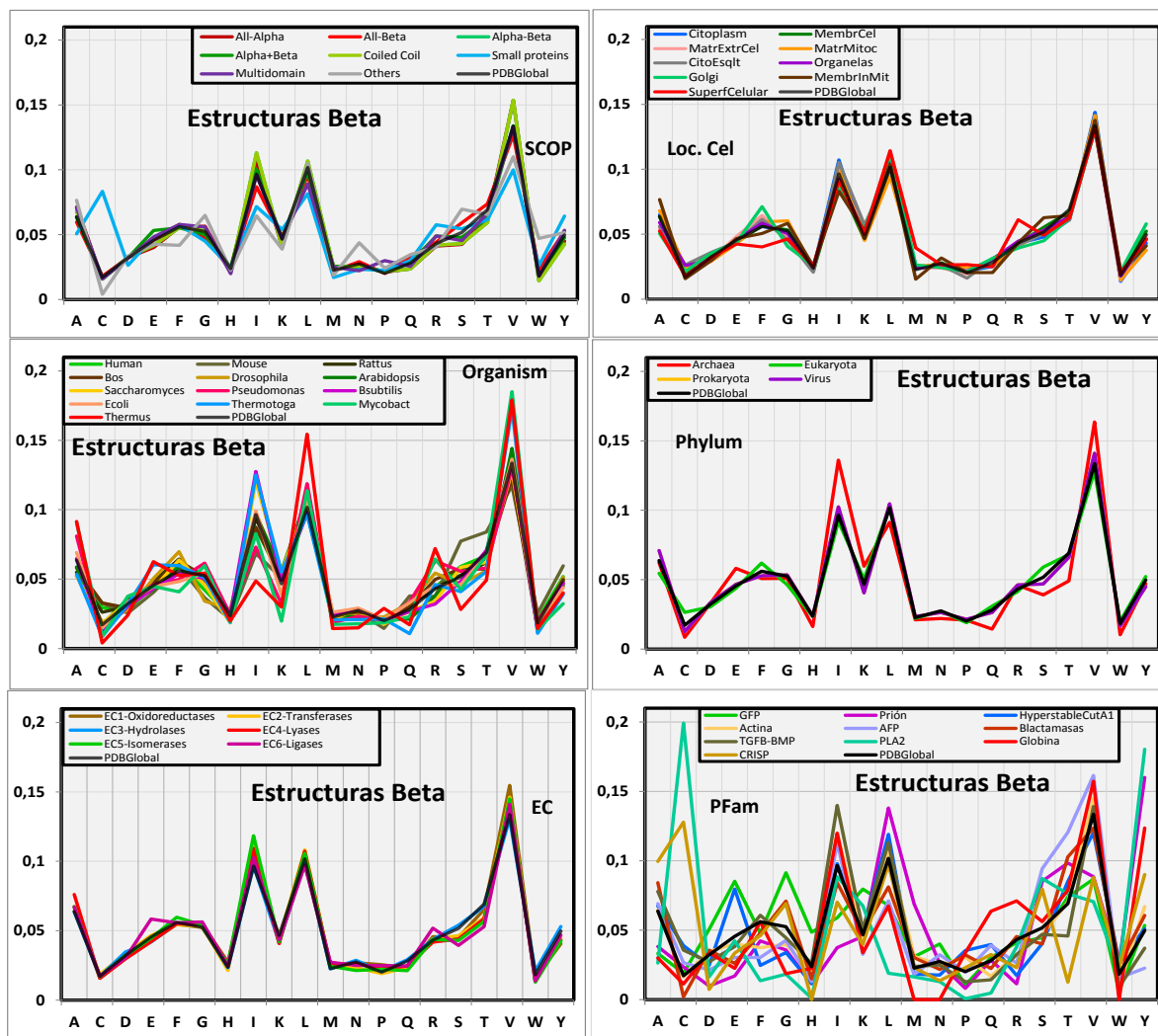


Figura 4.1.4.1. Frecuencia de aparición de aminoácidos en hilos y hojas Beta, en las proteínas analizadas en el presente estudio.

En las proteínas “otras”, la Cisteína tiende poco a aparecer en hilos/hojas Beta. El comportamiento de los demás aminoácidos es aproximadamente similar en todas estas proteínas.

Proteínas agrupadas por su localización celular

De nuevo, las mayores diferencias, se encuentran en proteínas de la Superficie celular, y en menor medida, en el Complejo de Golgi. Sin embargo, las diferencias no son muy grandes. Sólo vale mencionar que la Arginina en proteínas de superficie celular muestra la mayor frecuencia en hilos Beta, que las Argininas de las demás agrupaciones de proteínas. Por otro lado, Fenilalanina

muestra una mayor frecuencia en proteínas del Golgi, hacia los hilos Beta, que las demás proteínas.

Proteínas agrupadas por organismo

En proteínas de Thermus, la Valina es el mayor formador de estructuras de hilos/hojas Beta. El segundo mayor formador de esta estructura en Thermus, son las Leucinas, que como se vio en el PDBGlobal, presenta una frecuencia alta tanto en hélices, como en hilos y hojas Beta. En estas proteínas, la Cisteína aparece menos en esta estructura secundaria que en las demás proteínas del grupo. En el capítulo anterior se discutió sobre la presencia de este aminoácido en bacterias.

A manera de comentario sobre las proteínas humanas, es en ellas donde la Lisina aparece menos formando hilos/hojas Beta.

Proteínas agrupadas por Phylum

Al igual que con las hélices, en hojas Beta, las proteínas de Archaea son las que muestran un comportamiento diferente en sus frecuencias. Pero al igual que en la mayoría de las proteínas, la Valina es su mayor formador de estas estructuras secundarias. En cambio, la Isoleucina es mayor formador de hilos/hojas Beta, que la Leucina, diferente al comportamiento en PDBGlobal.

Proteínas agrupadas por Tipo de Enzima

En las enzimas, sólo hay una pequeña diferencia por resaltar, y es la de el Glutámico como un mayor aportante a las estructuras secundarias, que en otros tipos de enzima y en el PDBGlobal.

Proteínas agrupadas por Familia proteica

De nuevo las familias proteicas forman el grupo más heterogéneo en el aporte de los diferentes aminoácidos a la estructura secundaria considerada. Como comentario muy llamativo, las Cisteínas son el aminoácido más abundante en las estructuras de hilos/hojas Beta, apareciendo en casi el 20% de las mismas en proteínas de Fosfolipasa A2 (PLA2). La Tirosina (Y) es también un residuo alto formador de esta estructura secundaria, constituyendo alrededor del 18% de las mismas. A manera de contraste, en las β -Lactamasas no hay Cisteínas formando estructuras de hoja Beta. Las Prolinas en PLA2, tampoco aparecen en hilos Beta.

Las proteínas secretoras ricas en Cisteína (CRISP), también muestran un alto contenido de Cisteína en sus hilos Beta, con cerca del 13%. En ellas, ni la Histidina, ni la Prolina aparecen en este tipo de estructura secundaria. Algo similar ocurre con las Globinas en que ni Metionina ni Asparagina se encuentran formando esta estructura secundaria. La Glicina forma hojas Beta con mayor frecuencia en GFP que lo que forma en proteínas de otras familias.

4.1.5. Frecuencias de aminoácidos en Giros Totales, de las proteínas evaluadas en el presente estudio.

Hemos llamado Giros totales, a la suma de los giros Beta, mas Dobles, mas los Giros al azar (“random coils”). Esto equivale a lo que Chou y Fasman consideraron en su artículo clásico, como “giros” (Chou y Fasman, 1974a,b,c; Chou y Fasman, 1977, Costantini, 2002). Otros autores habían ya clasificado diferencias en los giros, pero en general, se siguió agrupando estos en una sola categoría (Richardson, 1981).

En esta sección, presentamos las frecuencias de aminoácidos en la formación de giros totales en las diferentes proteínas. A continuación, se muestra la figura 3.1.4.1., que agrupa las frecuencias en giros totales.

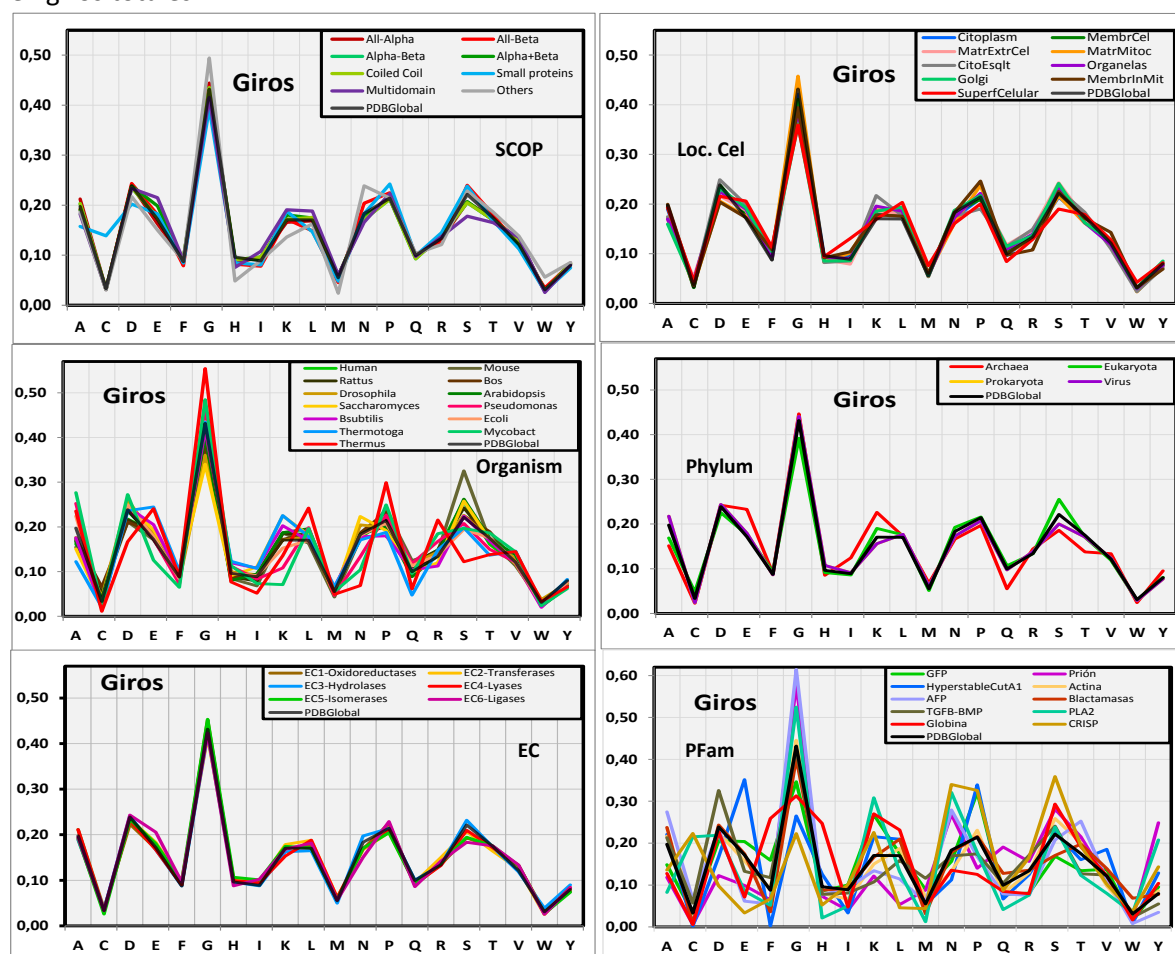


Figura 4.1.5.1. Frecuencia de aparición de aminoácidos en Giros totales, en las proteínas analizadas en el presente estudio.

Tendencias generales y Diferencias importantes

En la figura 3.1.5.1., se observa que los aminoácidos de las diferentes agrupaciones proteicas, son muy similares a las de la muestra PDBGlobal, en su tendencia a aparecer en estructuras de la

totalidad de los giros, presentando la mayor homogeneidad de las estructuras secundarias presentadas hasta ahora. Sin embargo, como se verá en las tres siguientes secciones, cuando se subclasifican los giros, aparecen diferencias importantes entre diferentes grupos de proteínas.

Las proteínas con mayor diversidad en sus frecuencias de aminoácidos hacia la aparición en giros, son las proteínas agrupadas por familia proteica.

Dado que las diferencias son menos notorias en este tipo de estructura secundaria, en esta subsección, agruparemos los comentarios de todos los grupos de proteínas en uno solo.

Con pequeñas diferencias, la Glicina aparece en estos giros, en las proteínas “Otras” en mayor proporción, que en las demás proteínas. En la mayoría de las proteínas de este grupo, la Cisteína presenta una baja frecuencia en giros totales, pero en las proteínas “pequeñas”, este aminoácido aparece con una frecuencia especialmente alta.

Los giros totales son predominantemente formados por el aminoácido Glicina y menos formados, según su frecuencia, por los residuos de Cisteína y de Triptófano. Sólo en las familias de proteínas aparecen otros aminoácidos con altas frecuencias en giros, como el caso del Aspártico como formador de giros en Proteínas Morfogenéticas de hueso (TGFB-BMP), el Glutámico y la Prolina en las proteínas Hiperestables CutA1, la Lisina y la Asparragina en las PLA2, La Asparragina, la Prolina y la Serina en las proteínas CRISP, y la Tirosina en las proteínas del Prión.

Esto refuerza nuestra idea de que las particularidades del microambiente en que actúa una proteína y la función que realiza, hace que se usen diferentes aminoácidos para la construcción de la misma estructura secundaria, para darle características particulares.

4.1.6. Frecuencias de aminoácidos en Giros tipo Dobléz, de las proteínas evaluadas en el presente estudio.

En la figura 3.1.6.1., se puede ver que los aminoácidos de las diferentes agrupaciones proteicas, son muy similares a las de la muestra PDBGlobal, lo que sugiere que en estas proteínas, los aminoácidos se comportan aproximadamente igual, en frecuencia de aparición en estructuras de dobleces. Por ello, las observaciones siguientes, no se subdividen en los grupos de proteínas.

En todos los grupos de proteínas, el aminoácido más frecuentemente encontrado en las estructuras de dobléz, son las Glicinas, y los tres residuos menos frecuentes en esta estructura secundaria, son Cisteína, Metionina, y Triptófano, con excepciones a esto en las proteínas agrupadas por familia proteica.

En las proteínas agrupadas por contenido de estructura secundaria (SCOP), localización celular, Phylum y tipo de enzima, se presenta poca variabilidad, con pequeñas excepciones: En las

proteínas clasificadas como “otras”, la Arginina (N), y la Treonina (T), presentan frecuencias más altas en dobleces que lo que presentan otras proteínas. En proteínas agrupadas por localización celular, las proteínas de superficie celular muestran mayor frecuencia del Glutámico formando dobleces que en otras proteínas. En proteínas agrupadas por Phylum, las proteínas de Archaea, son, al igual que en las estructuras secundarias anteriores, las más diferentes a las proteínas del PDBGlobal. Es de recordar que entre los organismos que integran a Archaea, hay varios grupos de organismos extremófilos, adaptados a ambientes con condiciones en que otros organismos no sobreviven, lo que implica que sus proteínas deben así mismo, presentar las adaptaciones adecuadas para conservar la estabilidad estructural a pesar de condiciones extremas de temperatura, salinidad, acidez y alcalinidad, por ejemplo. En proteínas agrupadas por tipo de enzima no se observan diferencias importantes.

En proteínas agrupadas por familia, la Glicina aparece casi dos veces más frecuentemente en estos tipos de giros, que en PDBGlobal. El aminoácido más frecuente en dobleces tanto en proteínas Hiperestables CutA1, como en GFP y en CRISP, es la Prolina. En Proteínas Morfogenéticas de hueso, es el Glutámico el más frecuente en esta estructura secundaria. En las Globinas, el aminoácido más formador de este tipo de giros, es la Fenilalanina. En PLA2, la Cisteína presenta una mayor frecuencia en giros que en ninguna otra familia de proteínas.

El residuo que más frecuentemente aparece en giros tipo dobléz, es la Glicina, seguida por Aspártico y Serina. Los menos formadores de estas estructuras, son La Cisteína, el Triptófano y la Metionina.

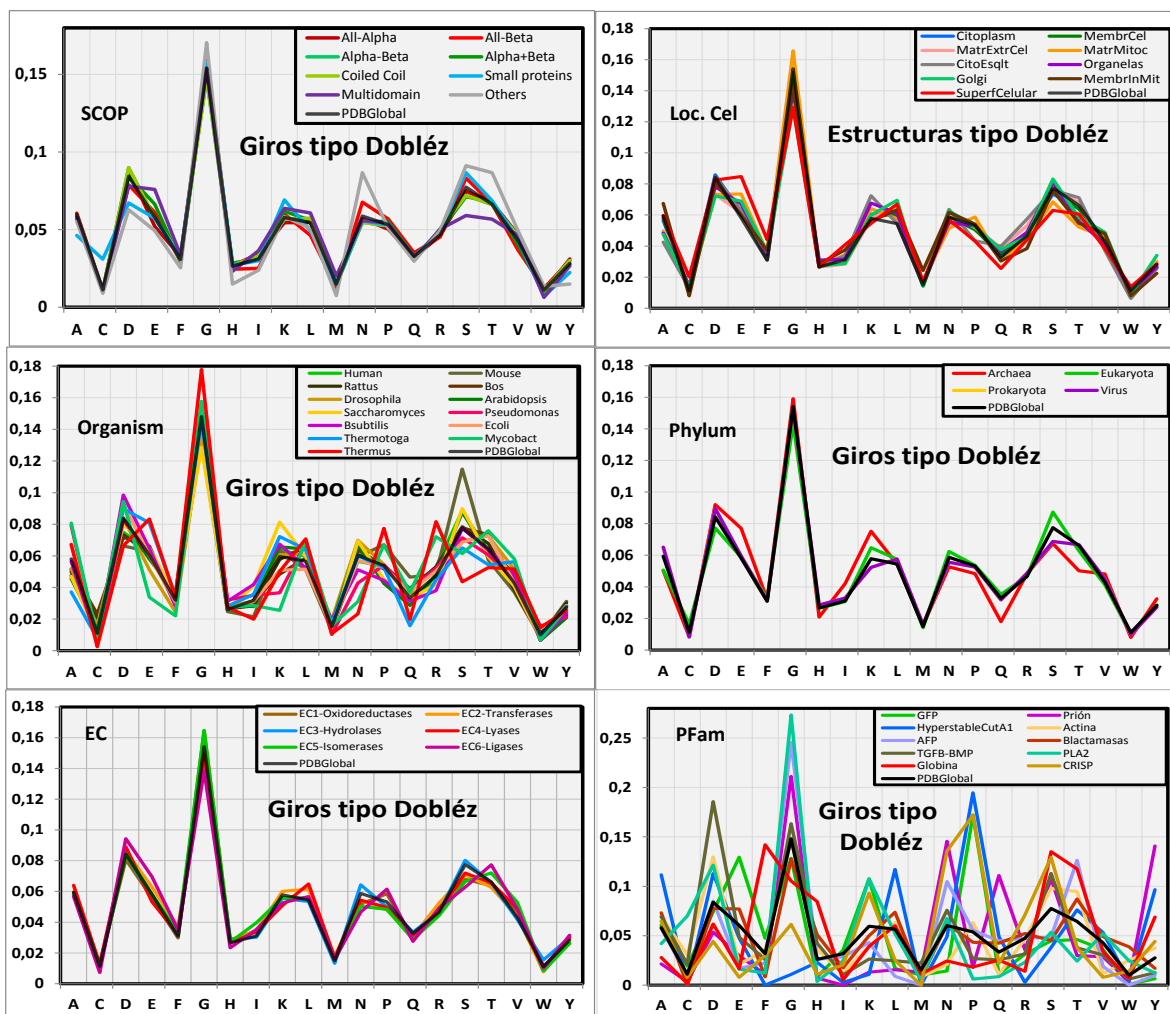


Figura 4.1.6.1. Frecuencia de aparición de aminoácidos en Giros tipo dobléz, en las proteínas analizadas en el presente estudio.

4.1.7. Frecuencias de aminoácidos en Giros tipo Beta, de las proteínas evaluadas en el presente estudio

En esta subsección se revisan las frecuencias de residuos en Giros tipo Beta en PDBGlobal, y los demás grupos de proteínas.

Tendencias generales y Diferencias importantes

En la figura 3.1.7.1. se pueden ver pocas diferencias en la mayoría de los cuatro grupos de proteínas menos variables en sus frecuencias de residuos en estructuras secundarias: en los grupos SCOP, localización celular, Phylum y Enzymas, la diferencias son mínimas; entre los diferentes organismos la variación, aunque un poco mayor, sigue pequeña.. En cuanto a familias proteicas aparece la mayor variabilidad. Lo más conspicuo, es que el aminoácido Glicina es el que

aparece más frecuentemente en giros Beta, con la excepción evidente de las proteínas Hiperestables CutA1, en las cuales el aminoácido más frecuente en giros Beta es el Glutámico. A manera de contraste llamativo, encontramos que el aminoácido más parecido al Glutámico, el Aspártico, no aparece en estructuras de giros tipo Beta, como tampoco la Cisteína aparece en ellos. En estas proteínas, otros aminoácidos presentan frecuencia cero en giros Beta; son ellos, Fenilalanina, Isoleucina, Metionina y Triptófano.

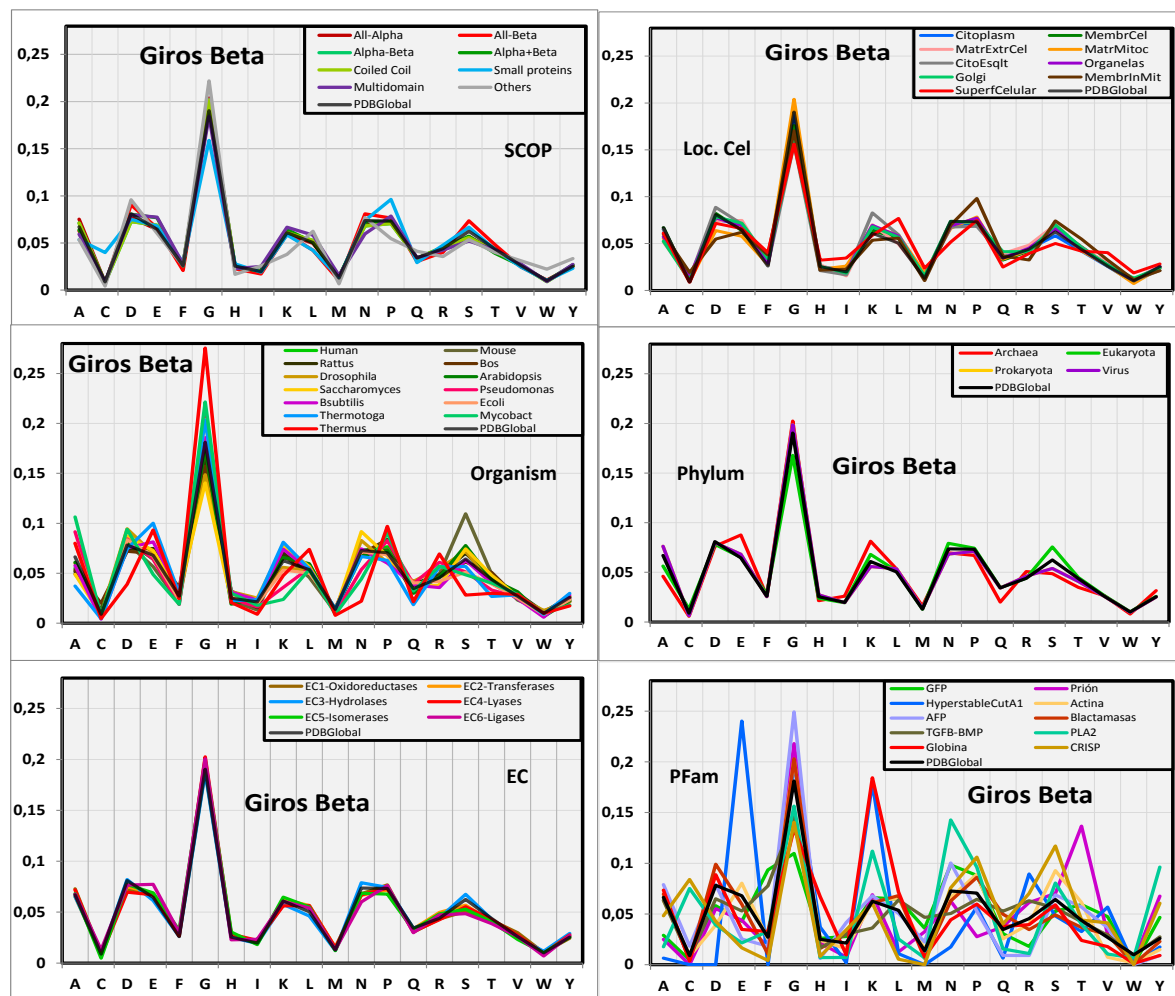


Figura 4.1.7.1. Frecuencia de aparición de aminoácidos en Giros tipo Beta, en las proteínas analizadas en el presente estudio.

En los priones, la Treonina también forma giros Beta, mientras que en las demás proteínas, aparece en estas estructuras, en frecuencias muy bajas. En CRISP, la Prolina y Serina también aparecen con frecuencia en estos tipos de giros.

Los residuos que más frecuentemente aparecen en estructuras de giros Beta, son las Glicinas, y no hay otro residuo con frecuencias tan altas. Los menos formadores de estas estructuras, son las Cisteínas, Triptófanos y Metioninas.

4.1.8. Frecuencias de aminoácidos en Random Coils “Giros al azar”, de las proteínas evaluadas en el presente estudio.

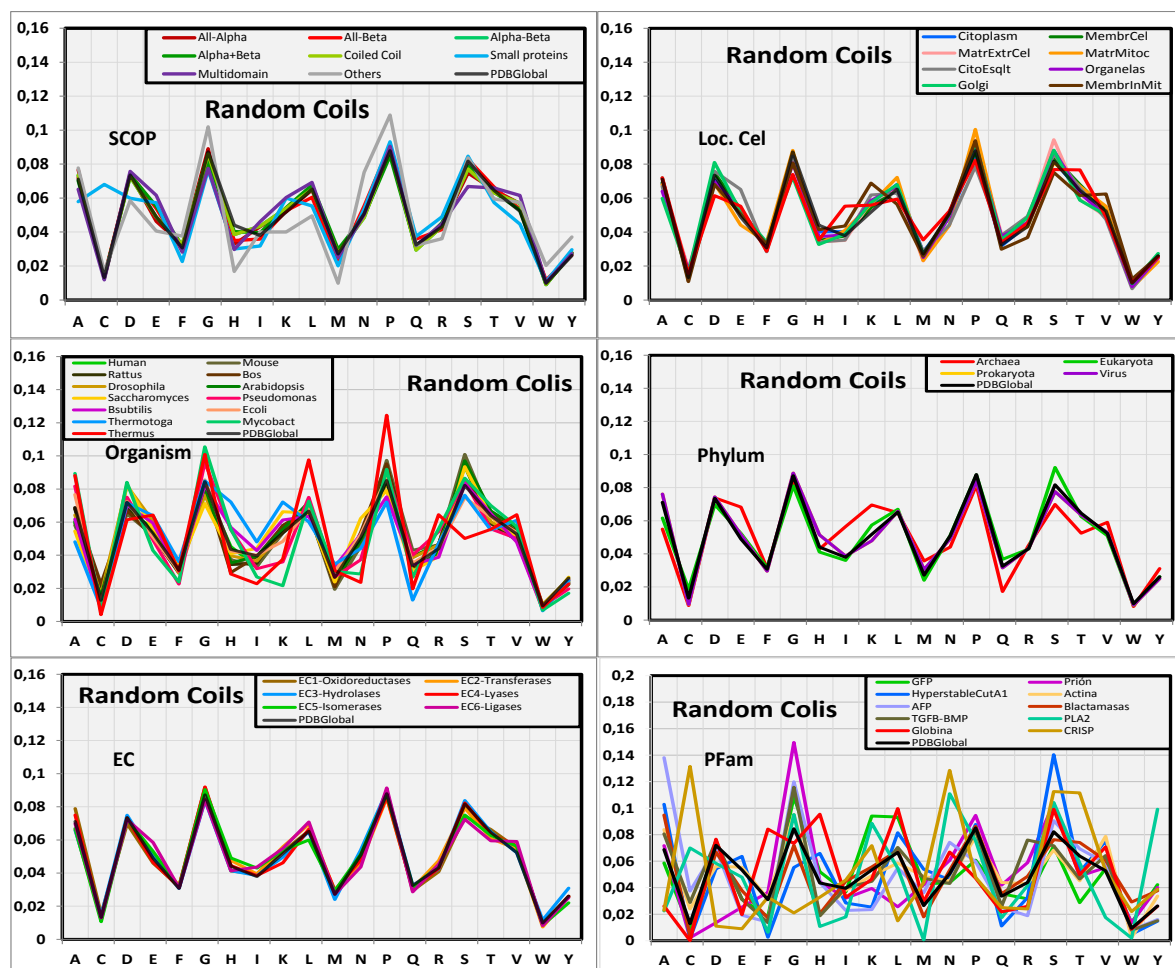


Figura 4.1.8.1. Frecuencia de aparición de aminoácidos en Giros al azar “Random Coils”, en las proteínas analizadas en el presente estudio.

Tendencias generales y Diferencias importantes

Aquí se han considerado las frecuencias de cada aminoácido en las estructuras al azar. De nuevo, las variaciones son pequeñas en proteínas agrupadas por Phylum, localización celular y tipo de enzima. En este caso, se observa más variación en familias de proteínas y en proteínas de organismo diferente.

A diferencia de los otros tipos de giro, en que las Glicinas eran las de mayor frecuencia, las Prolinas aparecen con mayor frecuencia, que las Glicinas, con la salvedad de que estas últimas aparecen con una frecuencia ligeramente menor y son los segundos aminoácidos fuertes formadores de giros aleatorios. En menor medida, aparecen frecuentemente las Serinas, Alaninas, Aspártico y

Leucinas, en estas estructuras. Los aminoácidos Cisteína y Triptófano, son los que menos aparecen en estas estructuras secundarias.

En las proteínas clasificadas como “otras”, la Prolina es aún más dominantemente frecuente en estos giros, que lo que ocurre en otras proteínas. En términos de frecuencia absoluta, el residuo que presenta la mayor frecuencia en giros al azar, es la Glicina, en los priones, en que su frecuencia en esta estructura está cerca del 15%. Aparte de ello, la frecuencia más alta de Prolina en estos giros, está en las proteínas de *Thermus*, con algo más del 12%. En la otra bacteria termo estable, *Thermotoga*, el aminoácido más formador de giros randómicos, es la Glicina, con algo más del 10%.

En las proteínas agrupadas por familia, se evidencia que no sólo las Glicinas y Prolinas forman giros al azar, pues en proteínas anticongelantes (AFP), son las Alaninas, mientras que en las proteínas CRIPS, es la Cisteína, seguida cercanamente en frecuencia por Asparragina, Serina y Treonina. Es decir, en estas proteínas, adaptadas a bajas temperaturas, los giros al azar se construyen con residuos más polares que en las demás proteínas. La Asparragina también es el residuo más común en esta estructura en las Fosfolipasas A2, seguido por Tirosina y Glicina. Además de la Glicina, el residuo más formador de giros al azar en GFP, son las Lisinas y Leucinas.

4.2. Frecuencias de tripletas en las estructuras secundarias

Se presentan aquí, las frecuencias de las tripletas y la estructura secundaria en la que aparecen, en el PDBGlobal. En este caso, se han analizado sólo las frecuencias.

4.2.1. Frecuencias de tripletas en estructuras secundarias, en el PDBGlobal

En la sección actual, se estudiaron las frecuencias de aparición de tripletas en estructuras secundarias. En este caso, al hablar de tripletas de residuos, las tripletas estructurales, es decir la estructura secundaria en que aparece cada triplete de aminoácidos, se agruparon de la siguiente forma: tripletas puras, de la forma “XXX”, en que los tres residuos de aminoácidos de la triplete, están en la misma estructura secundaria. Ellas son HHH, BBB, SSS, TTT, UUU, para indicar hélices, hilos Beta, Dobletes, Giros Beta y Giros al azar, respectivamente.

Luego se agruparon tripletas que hemos llamado “tripletas transicionales”, de la forma “XYY” ó “XXY”. En este grupo, quedaron las tripletas de transición de hélice a dobléz: HHS y HSS; transición de hélice a giro Beta (HHT y HTT); transición de hélice a giro al azar (HHU y HUU); transición de hélice a hilos/hojas Beta (HHB y HBB), y así, sucesivamente.

Finalmente, una categoría agrupó las tripletas de la forma “XYZ”, ó “XYZ”, en que al menos un solo residuo de aminoácido estaría asignado a una estructura secundaria, lo cual no tiene sentido desde el punto de vista de la estructura, pues no hay una estructura secundaria, como una hélice, que tenga una longitud de un solo aminoácido. A estas tripletas, que no hemos incluido en el estudio, les hemos llamado, “tripletras de difícil interpretación biológica”.

Para presentar los resultados de estas frecuencias, hemos agrupado las transiciones posibles entre dos tipos de estructuras secundarias ó tripletas estructurales puras. Es decir, a manera de ejemplo, en la primera gráfica se presentan las tripletas HHH y sus transiciones hacia y desde dobléz: HSS+HHS, SHH+SSH, y por supuesto, SSS.

Para todas estas tripletas estructurales, hemos calculado la frecuencia de aparición de cada tripleta de aminoácidos, calculada como el número de veces que aparece cada tripleta de aminoácido en cada tripleta estructural, dividido entre el número total de veces que aparece la tripleta de aminoácidos. Hemos escogido un umbral de frecuencia de aparición para comentar sobre las tripletas de mayor frecuencia de aparición en las proteínas del PDBGlobal.

En este caso, y como demostró el estudio de grupos de proteínas, nos hemos circunscrito al análisis del PDBGlobal y no a los grupos de proteínas, dado que buscamos valores que contribuyan a identificar reglas en el universo proteico.

Una vez se identificaron las tripletas que presentan mayor frecuencia de aparición en una estructura secundaria, hemos presentado las gráficas de la frecuencia de aparición de las mismas en las diferentes estructuras secundarias, con el fin de identificar las tripletas que presentan una especie de frecuencia alta absoluta, que permita identificar combinaciones de tres aminoácidos que muestren una fuerte tendencia hacia la formación de estructuras secundarias particulares.

Sobra decir que las tripletas de transición entre una estructura secundaria y otra, se han incluido en esta sección. Las siguientes figuras se componen de dos gráficas, la primera de las cuales, la de arriba en la figura, muestra las tripletas que más frecuentemente aparecen en una estructura secundaria dictada por la tripleta estructural correspondiente a cada tripleta de aminoácidos. La segunda gráfica, en la parte baja de cada figura, muestra que las tripletas que se identifican con alta frecuencia en una estructura, presentan su máxima frecuencia, justamente en esa estructura y no en otra.

Sobre las hélices y dobleces, y las transiciones entre ellos

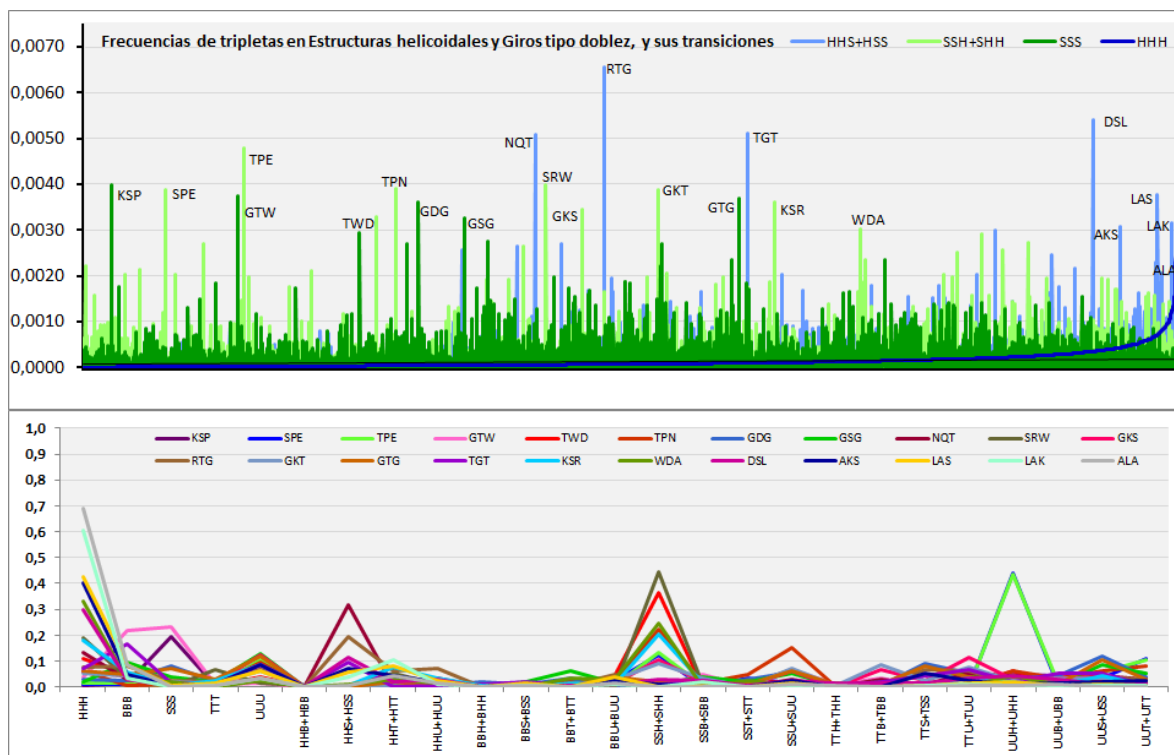


Figura 4.2.1.1. Tripletas con alta frecuencia hacia la formación de tripletas puras de hélices (HHH), dobleces (SSS), y las transiciones entre estas dos estructuras secundarias (HHS+HSS y SSH+SHH). En la gráfica de arriba, se presentan estas frecuencias. En la de abajo, las tripletas de mayor frecuencia.

Aunque en la parte de arriba de la figura se ven unas veintidos tripletas con altas frecuencias, en la gráfica de abajo, se evidencian sólo unas cuantas de ellas.

En la figura anterior, arriba, las tripletas más frecuentes en estructuras helicoidales (HHH), son ALA (Alanina-Leucina-Alanina), y AAL (Alanina-Alanina-Leucina), formadas por tres residuos no polares, y conocidos por su alta propensión Chou y Fasman hacia hélices. La segunda tripleta más frecuente en hélices, es LAK (Leucina Alanina Lisina). En la parte de debajo de la figura, en el extremo izquierdo, se ve que estas dos tripletas son las mayores formadoras de hélices en PDBGlobal.

Las tripletas, SRW (Serina Arginina Triptófano), y TWD (Treonina Triptofano Aspártico) tienen las frecuencias más altas en altas transiciones de dobléz a hélices (SSH+SHH). Como se puede ver, están constituidas por un residuo con cadena lateral hidroxilo, un triptófano y un residuo cargado.

La tripleta NQT (Asparagina-Glutamina-Treonina), es la tripleta con la mayor frecuencia en las transiciones entre dobleces y hélices, seguida por la tripleta RTG (Arginina-Treonina-Glicina).

La tripleta con la mayor frecuencia en la transición entre hélices y dobleces (SSH+SHH), es la tripleta NQT, formada por Asparagina-Glutamina-Treonina.

Sobre las Hélices y giros Beta, y las transiciones entre ellos

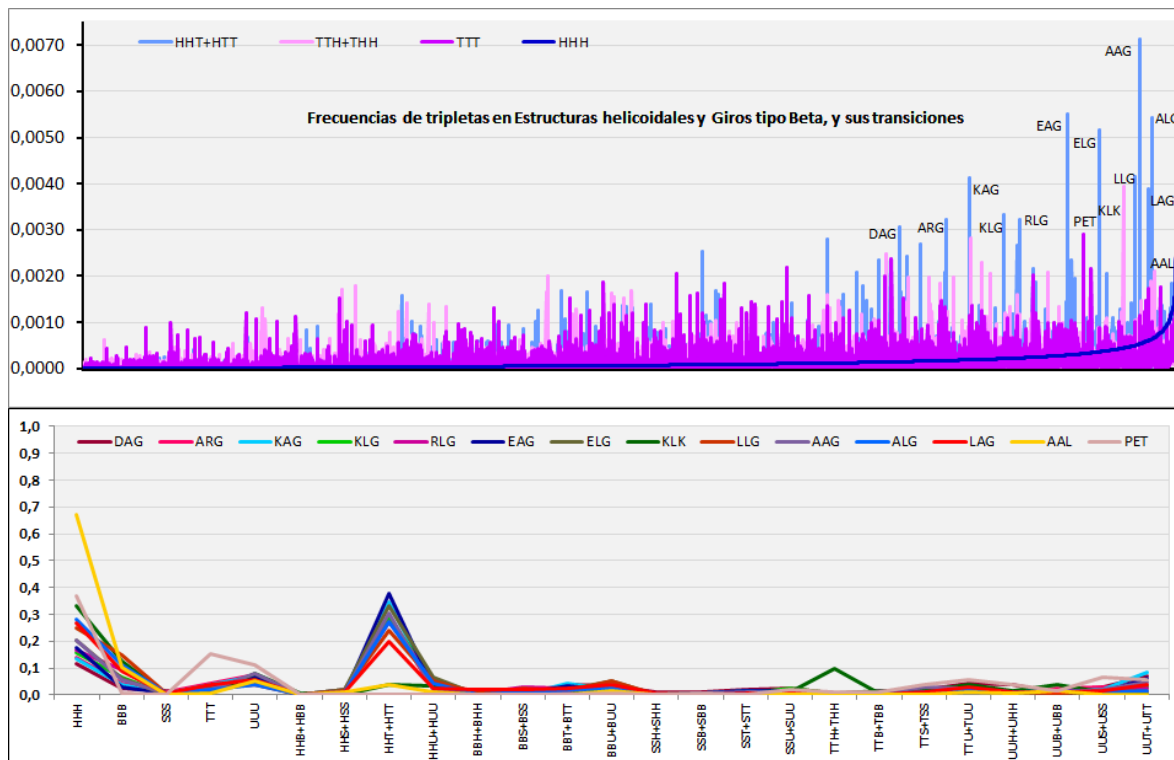


Figura 4.2.2.1. Tripletas con alta frecuencia hacia la formación de tripletas puras de hélices (HHH), Giros Beta (TTT), y las transiciones entre ellos (HHT+HTT y TTH+THH). En la gráfica de arriba, se presentan estas frecuencias. En la de abajo, las tripletas de mayor frecuencia.

Ya se comentó en la sección 3.2.1., que la tripleta AAL aparece con alta frecuencia en hélices.

Las tripletas con frecuencia más alta para la transición entre hélices y giros Beta, son: AAG, EAG, ELG, ALG, LLG, y otras, que en general, implican la terminación _AG ó _LG. Es decir, la combinación LG aparece frecuentemente en la transición de hélice e inicio de giros Beta. También se da la combinación _DG como iniciador de giros a partir de hélices y de otras estructuras secundarias (secciones subsiguientes de este capítulo). Esto es coherente con el hallazgo de que las Glicinas aparecen usualmente en la terminación carboxilo terminal de las hélices (C-cap) (Preissner y Bork, 1991; Bork y Preissner, 1991; Doig y Baldwin, 1995).

La tripleta que más frecuentemente aparece en medio de giros Beta en la forma TTT, es la tripleta PET (Prolina-Glutámico-Treonina). Es un dato llamativo, ya que el residuo individual que más aparece en esta estructura secundaria, es la Glicina. Sin embargo, el hecho de que Glicina sea muy frecuente, pero no aparezca en la tripleta estructural TTT, sugiere que la Glicina sea más un iniciador de giros Beta, que un “continuador” de esa estructura.

En la transición entre la terminación de un giro Beta y el inicio de una hélice, la tripleta más frecuente es KKK (Lysina-Leucina-Lysina). Esto diferente de la idea de que los iniciadores de hélices (N-cap), sean los aminoácidos aceptores de enlaces de hidrógeno, como Asparragina, Aspártico, Serina y Treonina (Doig y Baldwin, 1995). En este caso, un residuo cargado positivamente y dador de enlace de hidrógeno inicia una hélice. En los estudios anteriores, se han estudiado los inicios y finales de estructuras helicoidales, sin discriminar las transiciones desde y hacia otras estructuras secundarias.

En la parte de debajo de la figura 3.2.2.1. se puede verificar que las tripletas mencionadas, son efectivamente las que más frecuentemente forman las estructuras secundarias presentadas.

Sobre las Hélices y giros al azar, y las transiciones entre ellos

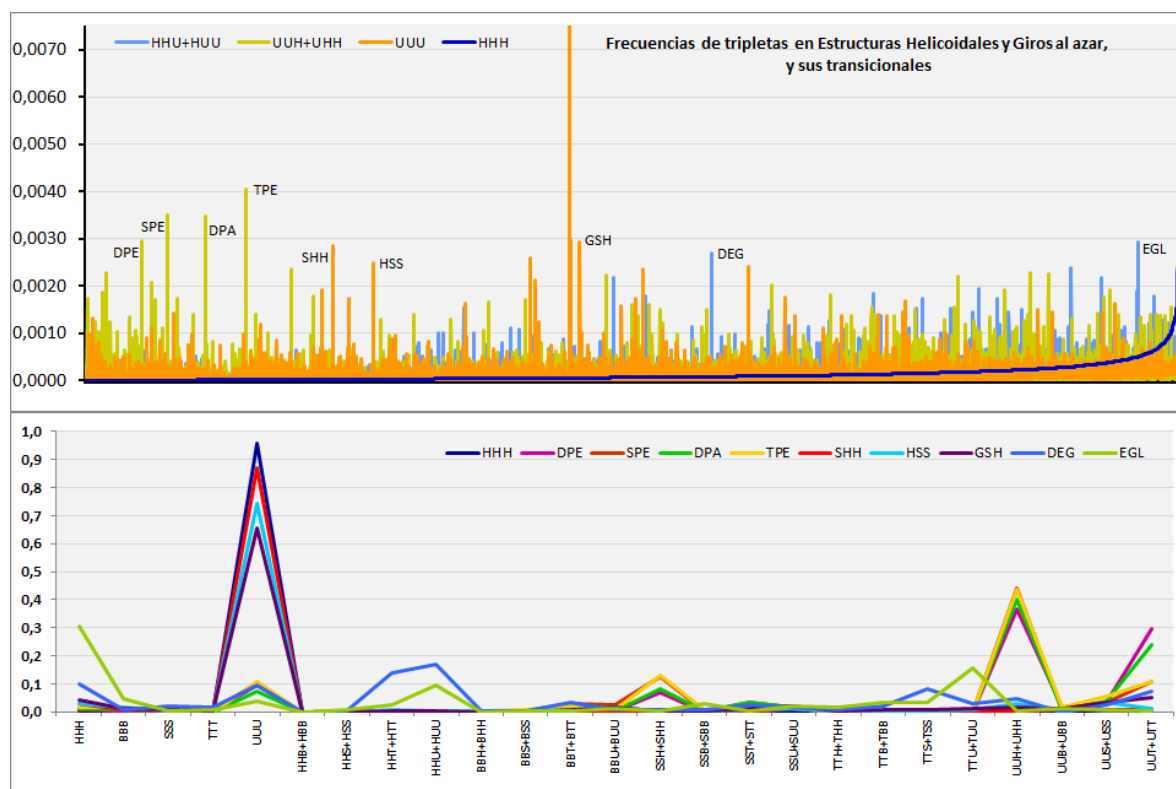


Figura 4.2.2.1. Tripletas con alta frecuencia hacia la formación de tripletas puras de hélices (HHH), Giros Beta (TTT), y las transiciones entre ellos (HHT+HTT y TTH+THH). En la gráfica de arriba, se presentan estas frecuencias. En la de abajo, las tripletas de mayor frecuencia.

En el capítulo anterior se habló sobre la presencia elevada de tripletas de Histidina (HHH) en el PDBGlobal. En este caso, se encuentra que la Histidina aparece con una frecuencia especialmente alta en estructuras de giros al azar, UUU (barra más alta, color naranja). Aparte de ella, la tripleta GSH (Glicina-Serina-Histidina). En la figura de abajo se puede ver que otras tripletas aparecen con mayor frecuencia en UUU, que son SSH (Serina-Histidina-Histidina), HSS (Histidina-Serina-Serina), y

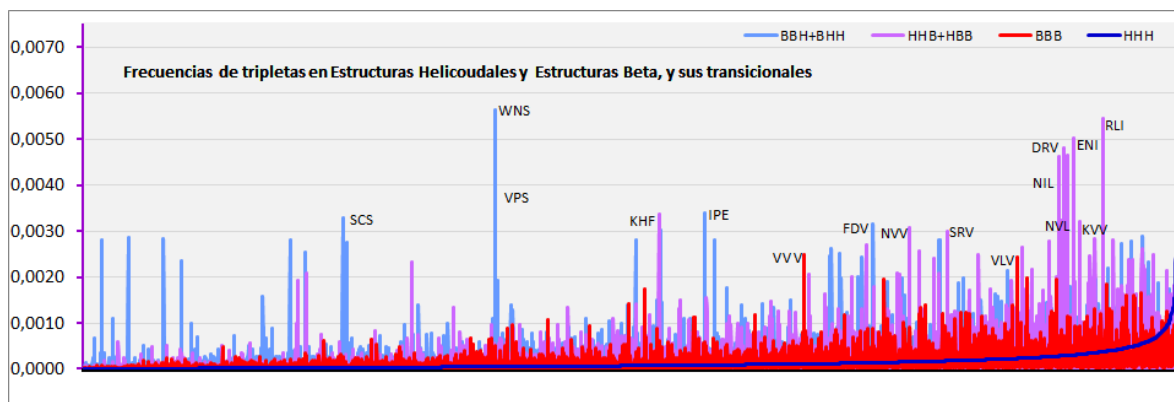
GSH (Glicina-Serina-Histidina), lo que muestra que las combinaciones con Histidina y Serina, pueden presentar una tendencia a la formación de estas estructuras aleatorias.

En cuanto a la transición entre hélices y giros al azar, las dos tripletas con la mayor frecuencia en ellas, son DEG (Aspártico-Glutámico-Glicina) y EGL (Glutámico-Glicina-Leucina). Al parecer la combinación de aminoácidos con cadena lateral ácido y Glicina, pueden contribuir a la transición de hélices a random coils.

Finalmente, la terminación de giros al azar para empezar estructuras helicoidales, encuentran en las tripletas TPE, SPE, DPE, DPA, y RPE (T:Treonina, S:Serina, P:Prolina, D: Ácido aspártico, E:Ácido Glutámico, R:Arginina). En este caso, las combinaciones de Prolina y un residuo ácido en el extremo carboxilo de la misma, parecen facilitar el paso de giros al azar, y hélices. Sobre decir que el Glutámico es un de los aminoácidos con mayor propensión hacia la formación de hélice, y que la Prolina es un “rompedor” de hélices, según Chou y Fasman (Chou y Fasman, 1974a,b,c; Chou y Fasman, 1977, Costantini, 2002).

Es importante anotar que en la totalidad de proteínas del PDB (PDBUniverso), las dos tripletas más abundantes fueron FQD y WWP, y que estas dos tripletas aparecen en estructura de giros al azar. Esto es importante, porque esta es una estructura importante dentro de la función proteica, pues aunque los giros al azar se constituyen en alrededor del 25% de las proteínas (al inicio de este capítulo), cerca del 50% de la función proteica, se encuentra en zonas de giros al azar (Tokuriki et al, 2008)

Sobre las Hélices y giros Beta, y las transiciones entre ellos



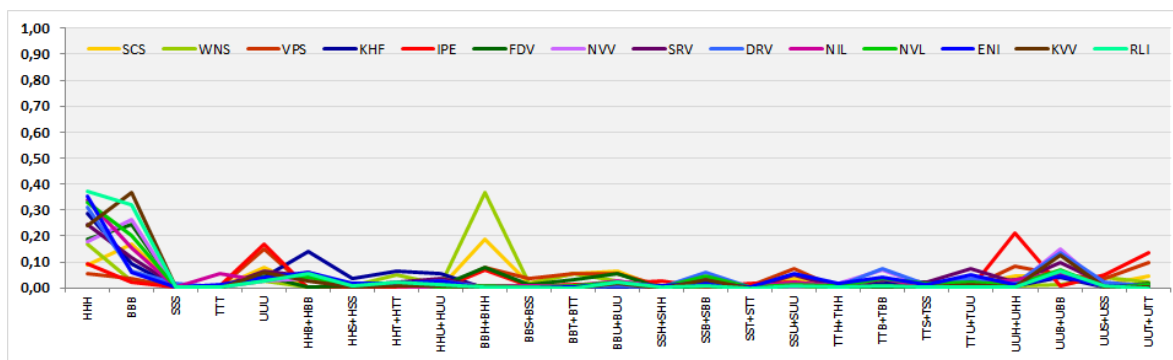


Figura 4.2.2.1. Tripletas con alta frecuencia hacia la formación de tripletas puras de hélices (HHH), Giros Beta (TTT), y las transiciones entre ellos (HHT+HTT y TTH+THH).

Como se vio en la primera sección de este capítulo 3, la Valina (V), es el residuo que aparece con mayor frecuencia en hilos/hojas Beta. Además, es el residuo con mayor propensión hacia esta estructura secundaria, según se ve en el siguiente capítulo 4, sección 1, y en la literatura (Nowick y Shabana, 1997). Pues bien, no es de extrañar, que la tripleta de Valinas (VVV), sea la más frecuentemente encontrada en estas estructuras. La segunda tripleta más frecuente en hilos/hojas Beta, es la tripleta VLV (Valina-Leucina-Valina); la Leucina también presenta una alta frecuencia como aminoácido individual en esta estructura, y una alta propensión hacia la formación de la misma.

La tripleta más frecuente en la transición entre hilo/hoja Beta y hélice, es WNS, formada por tres residuos que no presentan altas propensiones hacia hélices u hojas. Esto es llamativo, pues se esperaría que residuos poco frecuentes en una estructura secundaria específica, formaran tripletas con baja frecuencia en cualquiera de las estructuras secundarias; sin embargo, la tripleta WNS muestra que las tripletas permiten ver una parte del panorama: la de la tendencia a aparecer en las transiciones entre estructuras secundarias. En el siguiente capítulo, en que se analizan las propensiones, se verá que el valor de propensión de algunas tripletas no está dado hacia la formación de una estructura pura, sino hacia la formación de tripletas de transición entre estructuras secundarias, extendiendo el concepto de propensión hacia otro aspecto de la estructura secundaria de las proteínas.

Las tripletas frecuentes en la transición entre hélices y hojas Beta, son RLI (Arginina-Leucina-Isoleucina), ENI (Glutámico-Asparagina-Isoleucina), DRV (Aspártico-Arginina-Valina) y NIL (Asparagina-Isoleucina-Leucina). En general, combinaciones de residuos que contienen cargas negativas y positivas, combinados con aminoácidos no polares.

Sobre los giros tipo dobléz y los hilos/hojas Beta, y transiciones entre ellos

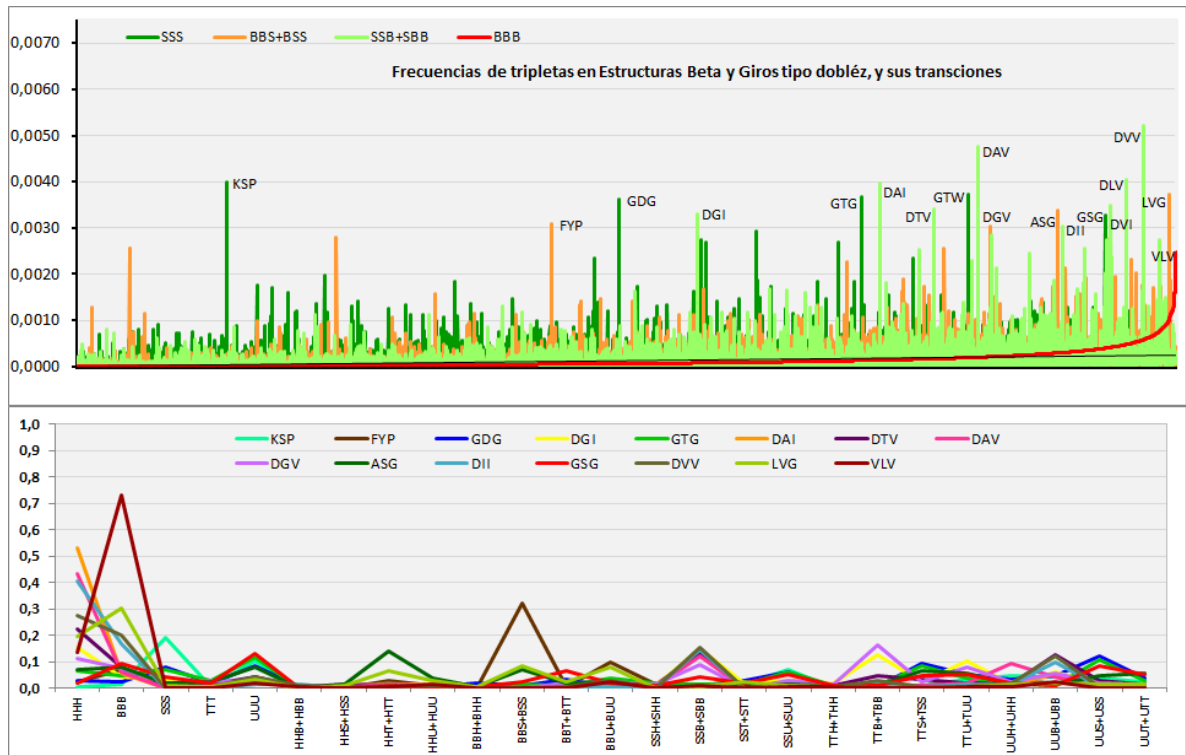


Figura 4.2.2.2. Tripletas con alta frecuencia hacia la formación de tripletas puras de hilos Beta (BBB), dobléz (SSS), y las transiciones entre ellos (BBS+BSS y SSB+SBB).

En la gráfica anterior, 3.2.2.2., se puede ver que las tripletas con mayor frecuencia en la estructura de giro tipo dobléz, son KSP, CDG, GTG y GSW.

CONCLUSIONES CAPÍTULO 4 - FRECUENCIAS DE LAS SECUENCIAS CORTAS EN LAS ESTRUCTURAS SECUNDARIAS

El contenido de estructura secundaria varía entre grupos de proteínas, siendo las estructuras helicoidales, las más comunes, y la suma de los tres tipos de giros, las siguientes estructuras más comunes.

De entre los tipos de giros en proteínas, los más comunes, son los giros al azar (“random coils”), que aparecen con frecuencia aflorando a la superficie de las proteínas, y resultan importantes para la función de una proteína en términos de su interacción con otras moléculas. Esto justifica que se le considere separado de los demás giros.

En general, Triptófano y Cisteína, son los residuos con menor frecuencia en cada una de las estructuras secundarias, lo que es inferible de la frecuencia de estos residuos en las proteínas. El Triptófano, es un aminoácido grande y no polar, cuyo transporte en el citoplasma celular, resulta costoso energéticamente para la célula. Al parecer, este aminoácido está involucrado en la formación de conformaciones que por su restricción estérica, adoptan formas geométricas muy particulares. La Cisteína, a pesar de su baja frecuencia de aparición, resulta de especial importancia para la estabilización de la estructura tridimensional. Parece que su papel está en la generación de enlaces covalentes a larga distancia y no es importante en la conformación de hojas o hélices. La tripleta Triptófano-Triptófano-Prolina (WWP), que es una de las más frecuentes en las proteínas (ver Cap 3), aparece casi siempre formando una estructura de giro al azar. Esto se ve reforzado con que es la tripleta con el Indicador de Preferencia Estructural (IPE) más alto, lo que indica que la estructura secundaria formada por estos residuos es muy estable (ver Cap 5). Es posible que dado que los giros tienden a aflorar a la superficie de las proteínas, esta tripleta sea importante en la función proteica, posiblemente en la interacción hidrofóbica con otras proteínas.

La Alanina, Leucina y Ácido Glutámico, son los aminoácidos que más frecuentemente aparecen en estructuras helicoidales, mientras que Cisteína, Triptófano, Prolina e Histitina, son residuos que aparecen poco en las estructuras helicoidales, lo que refuerza el concepto de propensión de algunos aminoácidos hacia la formación de estructuras secundarias, que se analiza en el capítulo siguiente. Esto se ve reforzado con el IPE para las tripletas formadas por al menos dos de estos tres aminoácidos, en que los valores de IPE son altos para la formación de estructuras helicoidales, tipo HHH, lo cual habla de una alta estabilidad de estos residuos en la formación de hélices. Estos resultados encuentran su análogo en un trabajo previo, publicado por nosotros, en que se miden secuencias cortas de aminoácidos, de diferentes longitudes (Mossos, et al, 2014). La tripleta AAL es la tripleta que aparece con mayor frecuencia en estructuras helicoidales.

Los residuos de Valina son los que más aparecen en estructuras de hilos y hojas Beta. Como se verá en el capítulo siguiente, las tripletas de Valina-Valina-Valina, presentan un IPE de más de 0.5, en la formación de hojas Beta, que implica que la mayoría de las veces, domina en esta estructura secundaria, y las hace estables. Después de Valina, los residuos que más aparecen en estas estructuras, son los de la Isoleucina. Los residuos de Valina e Isoleucina, son bastante parecidos en la estructura de su cadena lateral, y ambos son residuos preferidos para la construcción de hilos y hojas Beta. Sin embargo, el residuo más parecido a la Isoleucina es la Leucina, que es en contraste, un fuerte formador de hélices. Esto nos lleva a hipotetizar que la Leucina pudo haberse mutado en la estructura de su cadena lateral, convirtiéndose en Isoleucina para facilitar la formación de hilos Beta, siendo las Leucinas, más antiguas que las Isoleucinas (Trifonov, 2000)

Los residuos de Glicina, aparecen preferencialmente en Giros tipo Beta, y en menor medida en Giros tipo dobléz, mientras que los residuos de Prolina son los que predominantemente aparecen en estructuras al azar. Los giros tipo Beta requieren más cercanía entre los aminoácidos cercanos al giro, y la cadena lateral más pequeña es la de la Glicina; los dobleces podrían considerarse un giro Beta sin un enlace de hidrógeno estabilizador. La Prolina con su torsión debida al amino secundario, estaría causando la formación de un giro particular, diferente a los demás (ver conclusiones del capítulo 3).

Los residuos de Metionina, Asparragina, Glutamina, y Triptófano presentan frecuencias de aparición bajas y aproximadamente iguales en todas las estructuras secundarias, lo que indica que se acomodan en una estructura, dependiendo del contexto (es decir, de sus aminoácidos vecinos, y no determinan por sí mismos, la presencia de una estructura secundaria en particular.

Las proteínas virales presentan menores contenidos de estructuras helicoidales, y mayores contenidos de estructuras Beta que las proteínas de los demás Phyla.

Las proteínas agrupadas por familia proteica y por contenido de estructura secundaria, presentan las mayores variaciones en estructura secundaria. En ellas, las estructuras secundarias están construidas con diferentes combinaciones de aminoácidos.

Hay tripletas que aparecen casi exclusivamente en una estructura secundaria en particular, como WMC, que aparece más del 96% de las veces en HHH, la triplete Histidina-Histidina-Histidina, aparece más del 90% de las veces en UUU, SHH aparece más del 80% de las veces en UUU, para citar unos ejemplos, que además presentan un IPE mayor de 0.7, lo que indica que para hacer predicción de estructura, no se pueden considerar únicamente los aminoácidos individuales, sino que es conveniente mirar secuencias cortas en diferentes combinaciones de aminoácidos, y además, hay que mirar el contexto de secuencia. Esto puede ser una de las causas de las bajas predictibilidades de algunos modelos predictivos, basados en aminoácidos individuales.

El motivo de secuencia WWP, que en el capítulo anterior aparece en una muy alta frecuencia, forma un tipo de giro particular, posiblemente debido al tamaño de la cadena lateral del

Triptófano, y la torsión causada por el grupo amino secundario de la Prolina. Este giro particular se constituye en un tipo de giro que no está considerado entre las estructuras secundarias, y que reportamos como una nueva estructura, sumada a los otros tipos de giros.

BIBLIOGRAFÍA CAPÍTULO 4 – FRECUENCIA DE SECUENCIAS CORTAS EN ESTRUCTURAS SECUNDARIAS

Bork P, Preissner R. **On alpha-helices terminated by glycine. 2. Recognition by sequence patterns.** Biochem Biophys Res Commun. 1991 Oct 31;180(2):666–672.

Chou K.C. (2000) **Prediction of tight turns and their types in proteins.** Anal Biochem 286: 1–16.

Chou, P.Y., y Fasman, G.D. (1974). **Prediction of protein conformation.** Biochemistry—USA 13:222-45.

Chou, P.Y., y Fasman, G.D.(1974). **Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins.** Biochemistry—USA 13:222-22. 1974.

Chou, PY y Fasman, GD. (1974). **Prediction of protein conformation.** Biochemistry, 13, 222-245. El artículo clásico de Chou y Fasman.

Chou, PY. (1989) En Fasman, GD. (ed.), **Prediction of protein structure and the principles of protein conformation : Prediction of protein structural classes from amino acid compositions.** New York, pp. 549-586.

Chou PY, Fasman GD. (1977). **Prediction of protein conformation: secondary structure.** Journal of Molecular Biology., 115, 135-175.

Costantini, S., Colonna, G., Facchiano, A.M. (2006). **Amino acid propensities for secondary structures are influenced by the protein structural class.** Biochemical and Biophysical Research Communications 342, 441-451.

Doig, A.J., Baldwin, R.L. (1995). **N- and C-capping preferences for all 20 amino acids in alpha-helical peptides.** Protein Sci. 1995 July; 4(7): 1325–1336.

Mejía, D.F., Tischer, I. (2007). **Propuesta de un indicador de la tendencia de una secuencia de residuos de aminoácidos a encontrarse en estructuras secundarias particulares.** Pags 124-130 del Libro : “The 2nd. International Seminar on Genomics, Proteomics, Bioinformatics and Systems Biology”, Casa Editorial de la Universidad del Cauca, Popayan. ISBN 978-958-9451-21-2.

Mossos, N., Mejia-Carmona, D.F., Tischer, I. (2014). **FS-Tree: Sequential Association Rules and First Applications to Protein Secondary Structure Analysis.** Advances in Computational Biology - Advances in Intelligent Systems and Computing. 232: 189-198.

Neves, M.T., Jonson, P.H., Petersen, S.B. (1999). **Amino acid neighbours and detailed information analysis of Cysteines in proteins.** Protein engineering. 12:535-548.

Nowick, J.S., Shabana, I. (1997). **The Propensities of Amino Acids To Form Parallel β -Sheets**. J. Am. Chem. Soc., 119 (45):10903–10908.

Preissner R, Bork P. (1991). **On alpha-helices terminated by glycine. 1. Identification of common structural features**. Biochem Biophys Res Commun. 180(2):660-5.

Richardson, J.S. (1981). **The anatomy and taxonomy of protein structure**. Adv. Protein Chem., 34, 167-339.

Tokuriki, N., Stricher, F., Serrano, L., Tawfik, D.S. (2008). **How Protein Stability and New Functions Trade Off**. PLoS Computational Biology. 4(2):1-7.

Trifonov, E.N. (2000). **Consensus temporal order of amino acids and evolution of the triplet code**. Gene. 261:139-151.

CAPÍTULO 5. TENDENCIA DE AMINOÁCIDOS A ADOPTAR ESTRUCTURAS SECUNDARIAS PARTICULARES

En este capítulo analizamos la tendencia de los aminoácidos y tripletas para aparecer en determinadas estructuras secundarias. Para expresar estas tendencias, usamos tres indicadores diferentes: Las propensiones definidas por Chou y Fasman, la Entropía Conformacional y el Indicador de Preferencia Estructural, que introdujimos en un trabajo anterior (Mejía y Tischer, 2007), y que aplicaremos en los presentes análisis.

En la sección 5.1 evaluamos la propensión de Chou y Fasman en la muestra PDBGlobal y en las agrupaciones consideradas a lo largo de este estudio. La sección 5.2 se amplía este análisis contrastando la Propensión con la Entropía Conformacional y el Indicador de Preferencia estructural. La sección 5.3 presenta el análisis de las Propensiones de Chou y Fasman en tripletas de residuos de aminoácidos.

5.1. Propensiones Chou y Fasman

Se presentan aquí las propensiones, entropías e IPE, y se comparan unas con otras, resaltando el uso del IPE como complemento de la propensión.

5.1.0. Propensiones de C&F en las proteínas de PDBGlobal.

En el presente capítulo, se calcularon los valores de propensión de Chou y Fasman (C&F) para la formación de estructuras secundarias hélice α , Hoja β y giros (Pa, Pb, Pc), de acuerdo con lo reportado por los autores originales (Chou y Fasman, 1974; Chou y Fasman, 1978a,b), en la muestra PDBGlobal, y los diferentes grupos proteicos del presente estudio, y se compararon con las propensiones de Chou y Fasman, actualizadas y publicadas por Costantini y colaboradores, con una muestra de más de 2000 proteínas del PDB (Costantini et al., 2006), que hemos llamado muestra PDBLit, y cuyas frecuencias y propensiones se presentan en la tabla 5.1.0.1.

El cálculo de las propensiones de Chou y Fasman, para un aminoácido X en una estructura secundaria Y, se realiza de la siguiente manera, según Chou y Fasman (1974):

$$\frac{\# \text{ de veces que aparece X en Y } \div \# \text{ total de aminoácidos en Y}}{\# \text{ total de apariciones de X } \div \# \text{ total de aminoácidos}}$$

Con estas propensiones se pretende conocer la tendencia intrínseca de cada aminoácido en cada estructura secundaria.

En nuestra muestra PDBGlobal, y las submuestras de proteínas agrupadas por diferentes criterios, analizamos no sólo las propensiones de Chou y Fasman para las tres estructuras secundarias

Hélice Alfa, Hojas Beta y Giros, sino que agrupamos los tres diferentes tipos de hélices en una sola categoría de estructura secundaria, y lo mismo hicimos para hilos y hojas Beta. Sin embargo, los giros, sí se discriminaron en Giros Beta, en que hay un enlace de hidrógeno que contribuye a darle rigidez y estabilidad; dobleces, en que no hay presente un enlace de hidrógeno, lo que le da más flexibilidad, y giros al azar (Smith et al, 1996), en que aparentemente no hay una periodicidad ni regularidad identificable (Kabsch y Sander, 1983). Con esto, ofrecemos una novedad en cuanto a la identificación de la tendencia a formar estructuras secundarias.

En las siguientes figuras, se han comparado los valores de propensión Chou y Fasman entre nuestra muestra PDBGlobal, y los valores actualizados a partir de PDBLit (Costantini et al, 2006), donde los criterios de selección fueron similares a los usados en el presente trabajo. Como se observa en la figura 5.1.0.1., las dos muestras son esencialmente iguales en términos de las frecuencias de los aminoácidos.

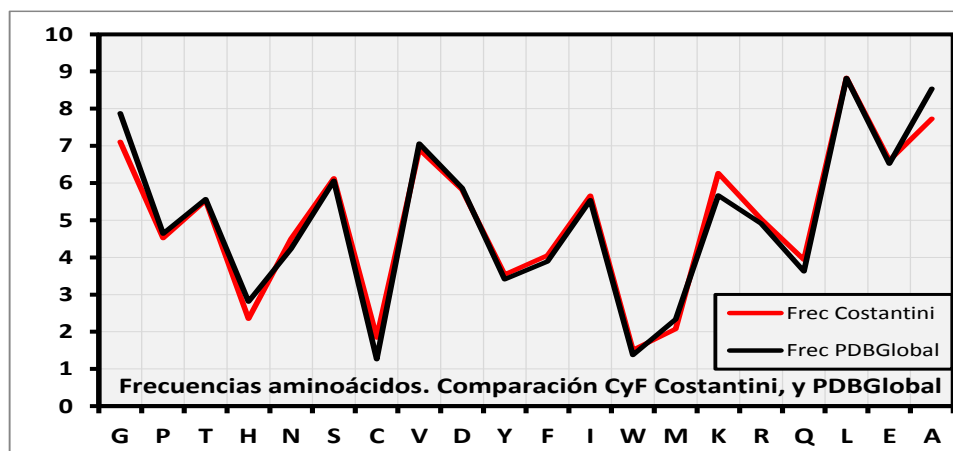


Figura 5.1.0.1. Comparación entre las frecuencias de los aminoácidos individuales entre la muestra PDBGlobal, y la muestra FrecLiterat, reportada por Costantini et al (2006).

En la siguiente tabla se presentan las propensiones clásicas de Chou y Fasman, con 15 proteínas, las actualizadas por Costantini et al (2006), con más de 2000 proteínas, y las calculadas en el presente trabajo, con más de 19000 proteínas. Pa, Pb y Pc, significan Propensión a formar hélices, Propensión a formar hilos y hojas Beta, y Pc: propensión a formar giros.

Tabla 5.1.0.1. Propensiones de Chou y Fasman, clásicos, calculados con una muestra de 15 proteínas (Chou y Fasman, 1974a), los calculados con una muestra de más de 2000 proteínas (PDBSelect) (Costantini et al., 2006), y los calculados en nuestra muestra de más de 19000 proteínas (PDBGlobal). Pa se refiere a la propensión a formar hélices; Pb, a formar hilos/hojas Beta, y Pc, a formar giros.

Aa	Pa-C&F	Pa-Const	Pa-PDBGlobal	Pb-C&F	Pb-Const	Pb-PDBGlobal	Pc-C&F	Pc-Const	Pc-PDBGlobal
A	1.41	1.39	1.42	0.72	0.75	0.75	0.82	0.80	0.79
C	0.66	0.74	0.85	1.40	1.31	1.36	0.54	1.05	0.93
D	0.99	0.89	0.91	0.39	0.55	0.55	1.24	1.33	1.32
E	1.59	1.35	1.39	0.52	0.72	0.70	1.01	0.86	0.84
F	1.16	1.01	1.00	1.33	1.43	1.44	0.59	0.76	0.76
G	0.43	0.47	0.47	0.58	0.65	0.67	1.77	1.62	1.61
H	1.01	0.92	0.76	0.80	0.99	0.84	0.81	1.07	1.28
I	1.09	1.04	1.03	1.67	1.71	1.74	0.47	0.59	0.57
K	1.23	1.11	1.15	0.69	0.83	0.82	1.07	1.00	0.98
L	1.34	1.32	1.30	1.22	1.10	1.15	0.57	0.68	0.67
M	1.30	1.21	1.14	1.14	0.99	0.97	0.52	0.83	0.90
N	0.76	0.77	0.77	0.48	0.62	0.65	1.34	1.39	1.38
P	0.34	0.50	0.55	0.31	0.44	0.44	1.32	1.72	1.67
Q	1.27	1.29	1.26	0.98	0.76	0.78	0.84	0.89	0.91
R	1.21	1.17	1.21	0.84	0.91	0.88	0.90	0.91	0.90
S	0.57	0.82	0.79	0.96	0.85	0.85	1.22	1.24	1.25
T	0.76	0.76	0.76	1.17	1.23	1.24	0.90	1.07	1.07
V	0.90	0.89	0.88	1.87	1.86	1.89	0.41	0.64	0.62
W	1.02	1.06	1.10	1.35	1.30	1.33	0.65	0.79	0.74
Y	0.74	0.95	0.97	1.45	1.50	1.46	0.76	0.78	0.77

5.1.0.1 Propensiones de Chou y Fasman para estructuras helicoidales

En la figura 5.1.0.2., se muestran los valores de propensión de Chou y Fasman, para estructuras helicoidales.

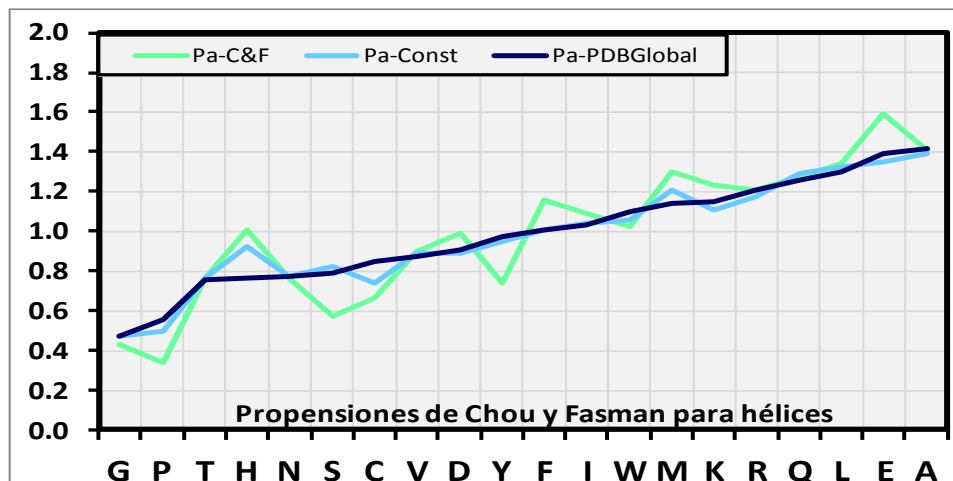


Figura 5.1.0.2. Propensiones de C&F, hacia estructuras helicoidales, para aminoácidos individuales, en C&F original, la muestra PDBLit por Costantini et al, y la muestra PDBGlobal.

Se observa que aunque las frecuencias de los 20 aminoácidos son muy similares, hay diferencias en las propensiones hacia la formación de hélices. Como puede observarse en la figura, las propensiones de Chou y Fasman, son las más diferentes, mientras que las de Costantini y las de PDBGlobal, están más cercanas entre ellas. Se observan pequeñas variaciones en Prolina (P), Cisteína (C) y Lysina (K), en que las de nuestra muestra son ligeramente mayores, e Histidina (H) y Metionina (M), menores en nuestra muestra con respecto a las propensiones actualizadas de

Costantini. Como es conocido y de esperarse, Glicina y Prolina son los residuos con menos tendencia a aparecer en hélices, mientras que Arginina (R), Glutamina (Q), Leucina (L), Glutámico (E) y Alanina (A), son los de mayor propensión hacia hélices.

5.1.0.2 Propensiones de Chou y Fasman para estructuras hojas y hilos

En la siguiente figura (5.1.0.3.), pueden verse las propensiones hacia la formación de hilos y hojas Beta obtenidas en PDBGlobal, en comparación con las obtenidas por Costantini en PDBLit. Como puede notarse allí, las propensiones hacia estructuras Beta son casi iguales, con la excepción notoria de la Histidina, en que el valor obtenido en PDBGlobal, es menor que en PDBLit.

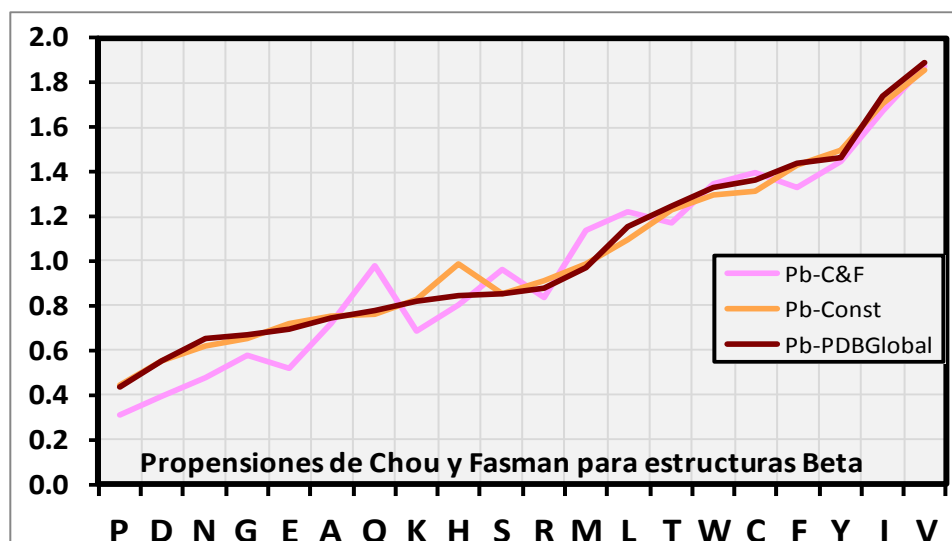


Figura 5.1.0.2. Propensiones de C&F, hacia hilos y hojas Beta, para aminoácidos individuales. Se presentan los valores originales de Chou y Fasman, la muestra PDBLit reportada por Costantini et al, y los de la muestra PDBGlobal.

En la figura 5.1.0.2., se ven diferencias muy pequeñas entre las propensiones actualizadas de Costantini y las de PDBGlobal, con la excepción notoria de la Histidina, que presenta mayor propensión en la muestra de Costantini. Es más notoria la diferencia con las propensiones C&F originales, con Prolina (P), Ácido Aspártico (D) y Asparagina (N) como los residuos menos formadores de hilos y hojas Beta, y algunos residuos de valores medios hacia estas estructuras.

5.1.0.3 Propensiones de Chou y Fasman para giros

En la siguiente figura (4.1.0.4), se pueden ver las propensiones para estructuras de giros (Pc), en las muestras PDBGlobal y PDBLit. Como puede notarse allí, aunque las propensiones son casi iguales en ambas muestras, hay pequeñas diferencias en los aminoácidos Triptófano (W) y Cisteína (C) que son menores en el PDBGlobal que en la muestra de Costantini, y los aminoácidos Metioinina

(M) e Histidina (H) que son algo mayores en propensión hacia giros en PDBGlobal que en PDBLit. La mayor diferencia está en Histidina que presenta casi 0.1 de diferencia en PDBGlobal, mayor que en PDBLit.

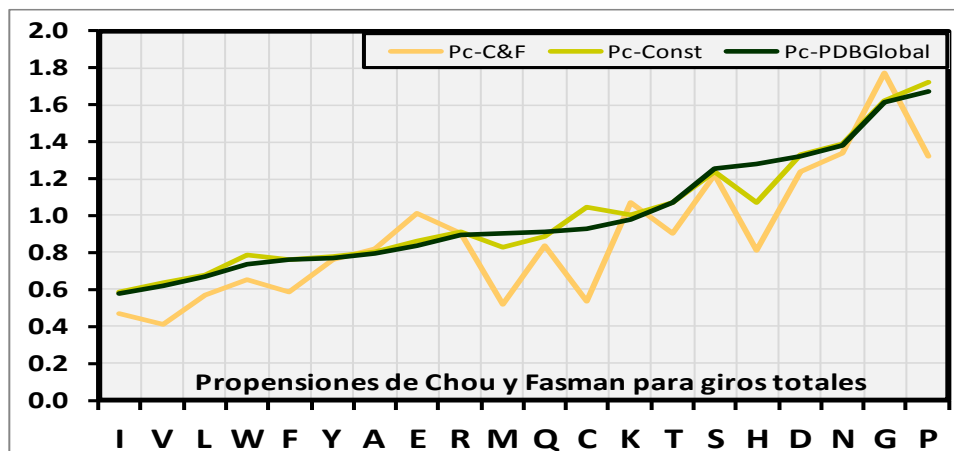


Figura 5.1.0.4. Propensiones hacia la formación de giros (Pc), calculados originalmente por C&F, y las de las muestras PDBGlobal y PDBLit.

En la figura anterior (4.1.0.4) es evidente que las propensiones actualizadas se diferencian de las propensiones originales, y que además, aunque hay mayor parecido entre las actualizadas por Costantini y las nuestras, hay diferencias mayores en los residuos Cisteína (C), e Histidina (H), y diferencias menores en los residuos Triptófano (W) y Metionina (M).

Los residuos menos propensos a aparecer en giros, son Isoleucina (I), Valina (V), Leucina (L), Triptófano (W), Fenilalanina (F), Tirosina y Alanina (A), es decir los residuos menos polares -más hidrofóbicos-, mientras que los residuos con mayor propensión hacia giros, son los residuos más hidrofílicos, lo que es coherente con la idea de que los giros tienden a aflorar a la superficie de las proteínas, interactuando con el entorno acuoso (Chou, 2000).

Por otro lado, es evidente que los residuos con mayor propensión hacia giros, son Glicina y Prolina, que además aparecen entre los de menos propensión hacia otros tipos de estructuras secundarias, como hélices e hilos Beta.

5.1.0.4. Propensiones de Chou y Fasman, subclasificando los tipos de giros en dobléz, giro Beta, y giro al azar

Dado que en el presente estudio hemos trabajado con las estructuras helicoidales (Alfa, Pi y 310) agrupada en una sola estructura de “helicoidales”, y los hilos Beta agrupados con las hojas Beta, las propensiones de Chou y Fasman para ellas, son esencialmente las mismas. Sin embargo, en el presente estudio, hemos subdividido los giros, en consideración a su importancia en la movilidad de las estructuras proteicas, a su importancia como conectores entre otras estructuras secundarias, y a su evidente participación en la función de las proteínas. En la Figura 5.5. se pueden ver las propensiones para los diferentes tipos de giros (doblez, giro Beta con enlace de H estabilizante y Giro al azar).

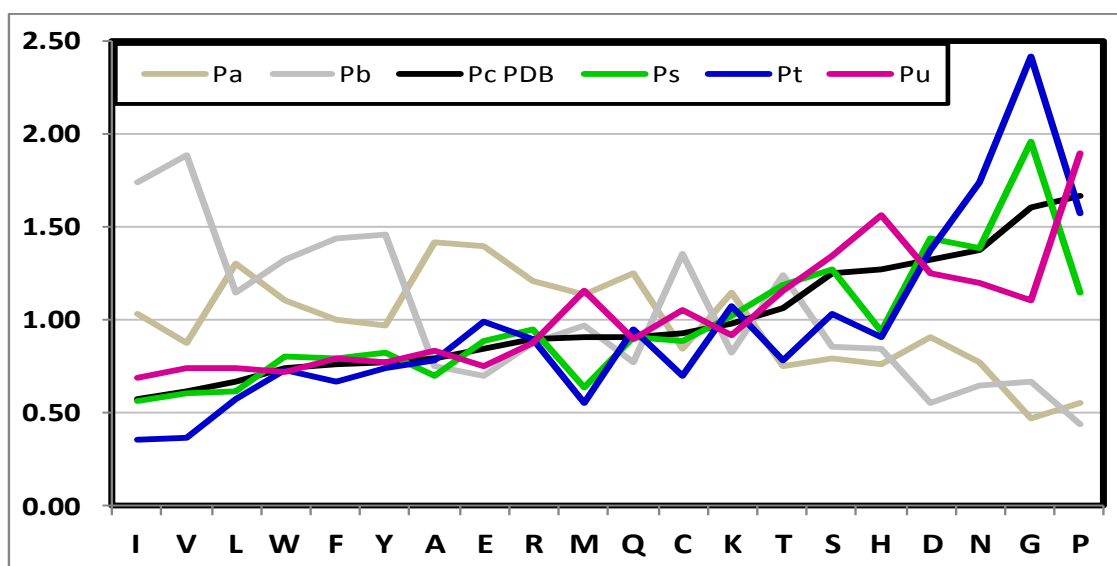


Figura 5.1.0.5. Propensiones de Chou y Fasman calculadas para los tres tipos de giros, según DSSP.

En la figura 5.1.0.5., las convenciones son: Ps=Propensiones a formar dobléz; Pt=Propensiones a formar giros Beta con un enlace de hidrógeno que los estabiliza, y Pu=Propensiones a formar giros al azar. En color negro, la propensión de Chou y Fasman calculada para los giros en general, si subdividirlos. En color gris, la propensión a formar estructuras Beta para cada aminoácido, y en color crema, la propensión a formar estructuras helicoidales.

Algunos aminoácidos presentan una mayor tendencia a formar unos tipos de giros, que otros. Por ejemplo, la Metionina presenta mayor tendencia a formar giros al azar, que otros tipos de giros, al igual que Cys. En las propensiones a formar giros al azar, los aminoácidos que presentan mayor tendencia a formar estos tipos de giros, son, en orden descendente: Prolina, Histidina, Treonina, Cisteína, Metionina e Isoleucina. La Prolina y la Histidina son fuertes formadores de giros, y por encima de sus propensiones hacia otras estructuras secundarias, tienden a formar estructuras al azar. También es notorio que en general, los residuos fuertes formadores de Giros, son los más

débiles formadores de estructuras helicoidales e hilos/hojas Beta, y viceversa, los fuertes formadores de hélices e hilos/hojas, son débiles formadores de giros.

La Glicina es el más fuerte formador giros Beta, y aunque en menor medida, también es un fuerte formador de dobleces, mientras que su propensión más baja entre los giros, es a formar estructuras al azar. La Prolina por su lado, es el máximo formador de giros al azar, y su menor propensión entre los giros, es a formar dobleces.

La Glicina presenta la mayor propensión hacia una estructura secundaria de entre todos los veinte aminoácidos, con una propensión de 2.42 hacia la formación de giros Beta, posiblemente dado el tamaño más pequeño de su cadena lateral, con lo que permite el mayor acercamiento entre las cadenas laterales de los aminoácidos vecinos en la cadena. El segundo aminoácido con mayor propensión a una estructura secundaria, es Valina, con una propensión de 1.89 hacia la formación de estructuras Beta,

Los aminoácidos Treonina (T), Serina (S), Histidina (H), Aspártico (D), Asparragina (N), Glicina (G) y Prolina (P), presentan mayores propensiones hacia la formación de giros que hacia la formación de estructuras beta ó estructuras helicoidales. Mientras que los aminoácidos con altas propensiones a hojas ó hélices (Ile, Val, Leu, Trp, Phe, y Tyr), tienen a presentar bajas propensiones hacia la formación de giros en proteínas.

También es evidente en la gráfica, que hay un grupo de aminoácidos cuyas propensiones tienden a no ser claramente fuertes hacia la formación de estructuras secundarias específicas, como Glutámico (E), Arginina (R), Metioina (M), Glutamina (Q), Cisteína (C), y Lisina (K).

5.1.1. Propensiones de C&F en proteínas agrupadas según SCOP.

En la presente sección, se analizarán las propensiones C&F en las proteínas agrupadas de acuerdo con su contenido de estructura secundaria, según SCOP.

Como puede observarse en la figura 5.1.1.1., las proteínas “otras”, presentan las mayores diferencias en las propensiones hacia estructuras de hilos/hojas Beta, y los diferentes tipos de giros, con las Cisteínas formando más fuertemente giros al azar que las demás proteínas de esta clasificación SCOP (Cisteína presenta una propensión de más del doble que las proteínas del PDBGlobal y que las demás proteínas de esta agrupación SCOP); mientras tanto, el mismo residuo en las proteínas “otras”, presenta una propensión hacia la formación de hilos/hojas Beta, de menos de la mitad que en las demás proteínas. En general, la Cisteína en este grupo de proteínas, presenta una tendencia intrínseca más baja hacia la formación de hélices, hojas, dobleces y giros Beta que en las demás proteínas.

A manera de contraste, en el mismo grupo de proteínas “otras”, el residuo de Metionina presenta una mayor propensión hacia hélices que hacia los hilos/hojas Beta, y marcadamente menor hacia la formación de los giros en general. En general, las proteínas “pequeñas” son las que presentan las mayores diferencias en sus propensiones hacia la formación de las diferentes estructuras secundarias, con la excepción de la tendencia a la formación de hélices, en que las proteínas “pequeñas” presentan las mayores diferencias.

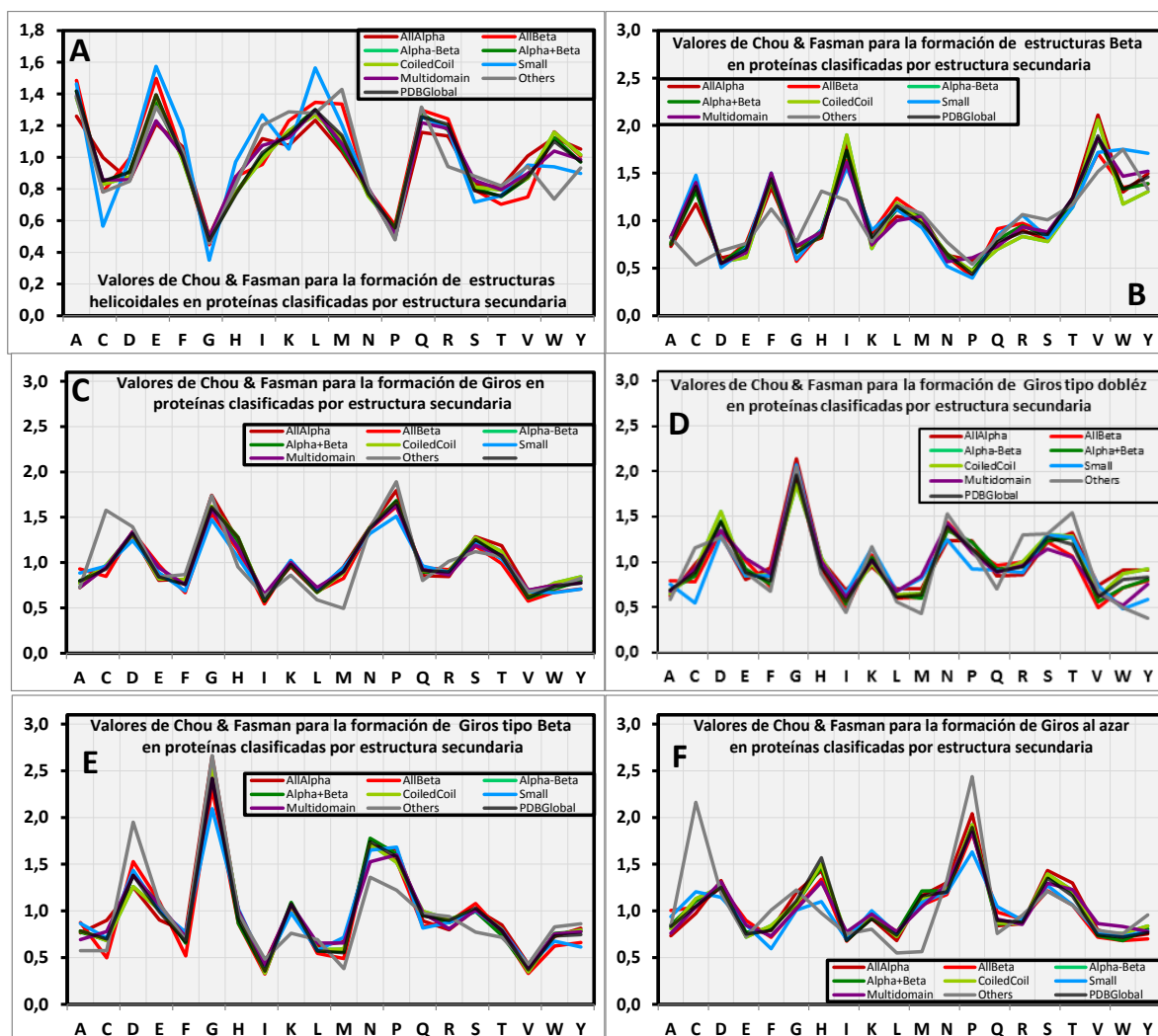


Figura 5.1.1.1. Propensiones de C&F, en proteínas agrupadas según SCOP.

Además de la diferencia en la Cisteína para la formación de giros al azar, la Prolina presenta en las proteínas “otras”, la máxima propensión hacia la formación de giros al azar. La Glicina es consistentemente alta en su tendencia intrínseca a la formación de dobleces y giros Beta.

Las diferencias entre algunos grupos de proteínas y el PDBGlobal, sugiere que la propensión de Chou y Fasman generalizada para todas las proteínas, puede estar desconociendo particularidades de grupos de proteínas, y por tanto podría contribuir a explicar la baja predictibilidad que en ocasiones genera el enfoque de Chou y Fasman.

5.1.2. Propensiones de C&F en las proteínas agrupadas por localización celular

En esta sección, se analizan las tendencias intrínsecas conocidas como Propensiones de Chou y Fasman, hacia la formación de las cinco estructuras secundarias (DSSP-5). Como puede observarse en la figura 5.1.2.1. parte A, las propensiones hacia la formación de estructuras helicoidales son bastante homogéneas entre las proteínas agrupadas por localización celular, con pequeñas diferencias en el residuo de Tirosina (Y) , que presenta una propensión mayor hacia hélices, en las proteínas de la superficie celular, que en PDBGlobal y en las demás proteínas de esta agrupación.

En las propensiones hacia la formación de hilos /hojas Beta (parte B de la figura 5.1.2.1.), las mayores diferencias se presentan en los residuos Leucina (L) y Metionina (M) de las proteínas de la matriz extracelular, que presentan mayores propensiones hacia esta estructura, que en las demás proteínas de este grupo. La Metionina y la Fenilalanina (F) en las proteínas de la superficie celular, tienden menos a formar hilos Beta, que las otras proteínas del grupo.

En cuanto a la formación de dobleces, las proteínas de la matriz extracelular muestran que la Cisteína en estas proteínas, presentan las propensiones más bajas.

En general, las proteínas cuyas propensiones hacia hélices e hilos/hojas Beta, se alejan más de las propensiones Chou y Fasman calculadas para el PDBGlobal y para las proteínas agrupadas por su localización celular, son las proteínas de la matriz extracelular.

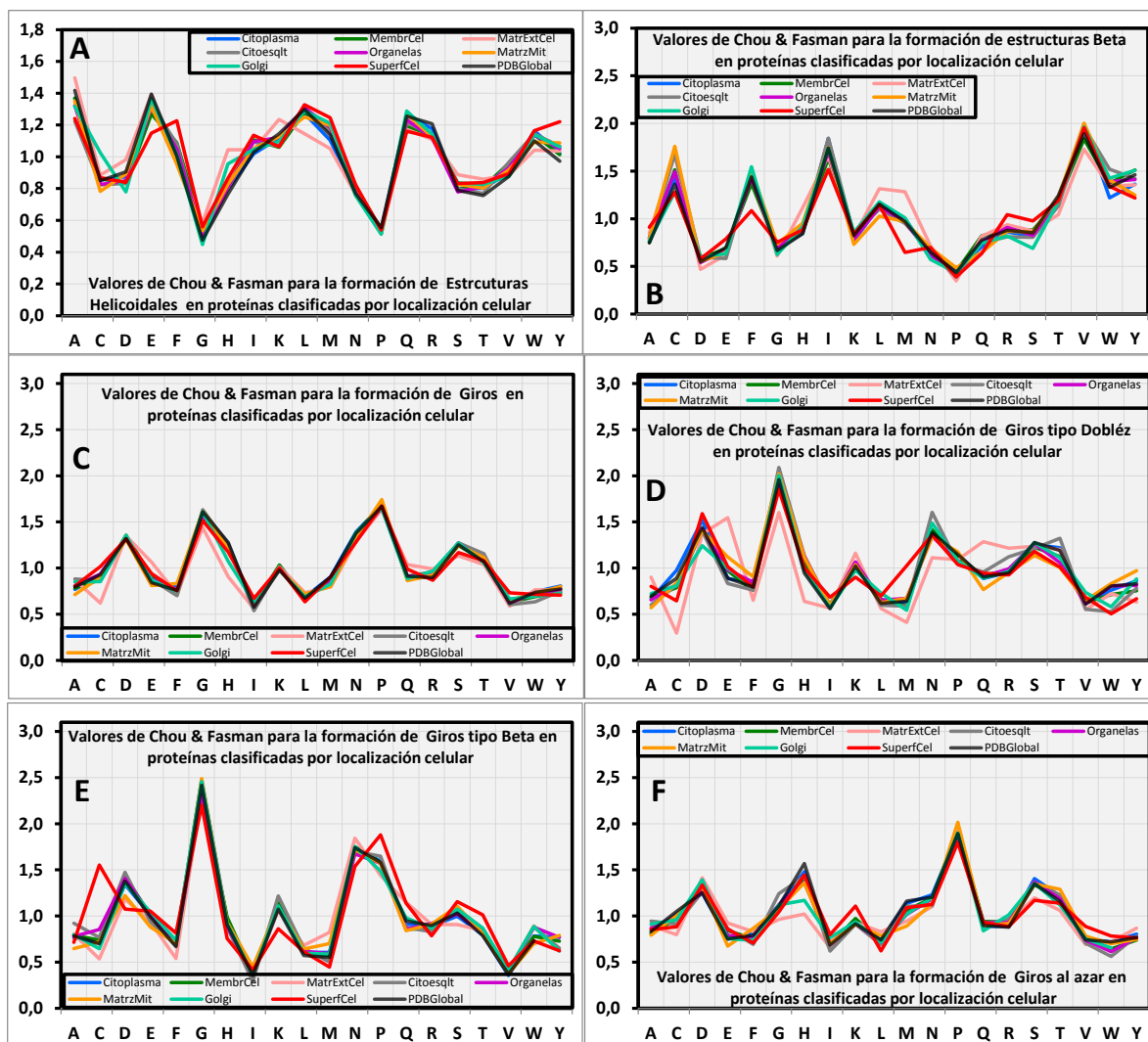


Figura 5.2.1. Chou & Fasman de aminoácidos en estructuras helicoidales, en proteínas agrupadas por su localización celular.

En la formación de giros, no se encontraron diferencias importantes entre los diferentes grupos de proteínas agrupados según SCOP, con la excepción de la Cisteína, como un aminoácido con menor propensión a la formación de giros que lo que el mismo residuo forma en las demás proteínas.

5.1.3. Propensiones de C&F proteínas agrupadas por organismo

A continuación, se presentan algunas curiosidades en las propensiones de los aminoácidos hacia estructuras secundarias, en proteínas agrupadas por el organismo del que provienen.

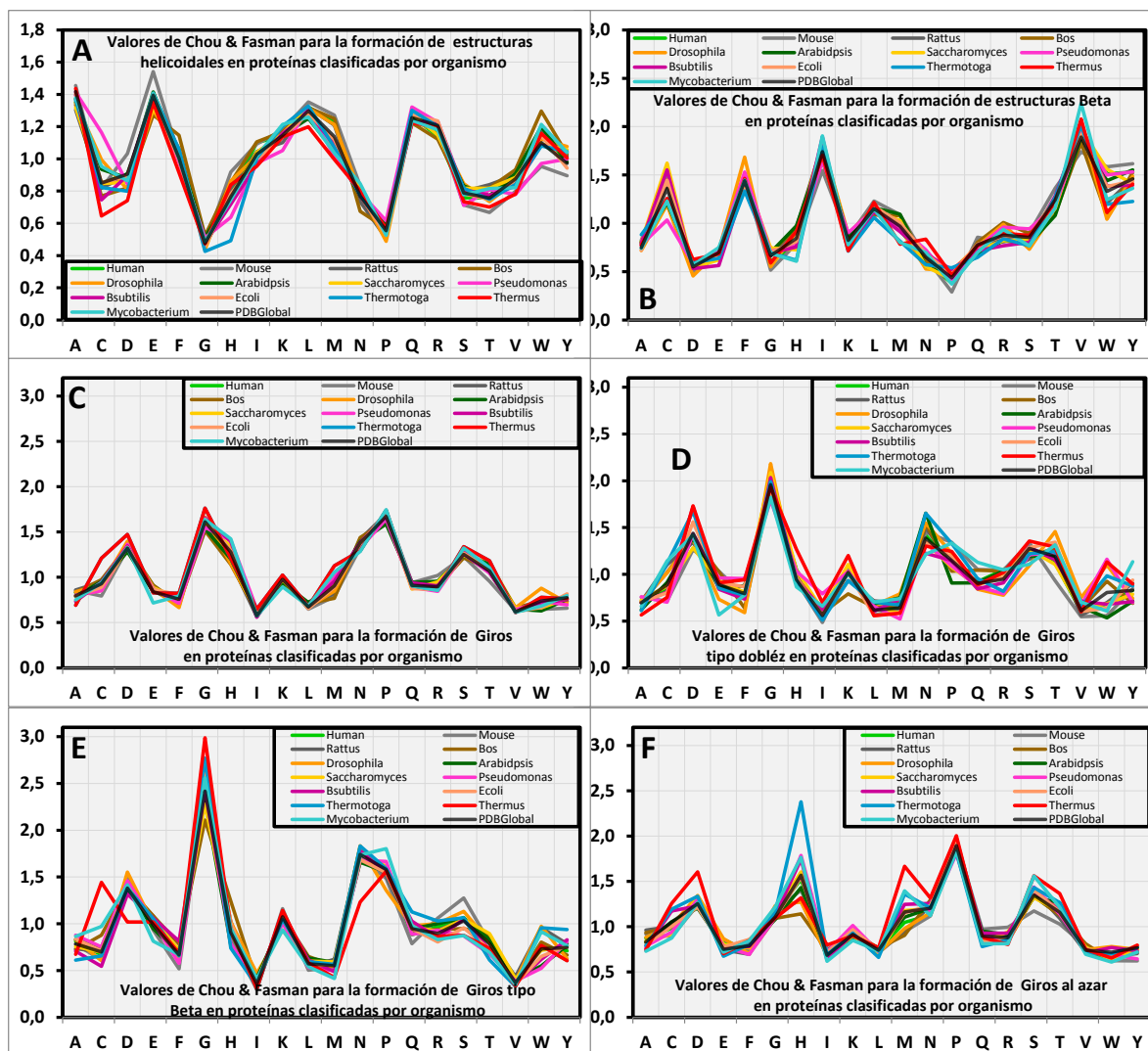


Figura 5.1.3.1. Propensiones Chou & Fasman de aminoácidos en proteínas agrupadas por Organismo.

Las menores variaciones en los valores de propensión, las presentan los hilos/hojas Beta. Hay pocas variaciones en las propensiones hacia estructuras helicoidales. Los giros vistos como generalidad, presentan propensiones muy homogéneas. Sin embargo, al mirar los giros más en detalle y discriminarlos según el tipo de giro, saltan a la vista otras diferencias llamativas, como que la Cisteína, de baja tendencia a formar giros Beta en la generalidad (PDBGGlobal), tiende a formar estas estructuras secundarias más de dos veces más altamente en las proteínas de *Thermus* que en la generalidad de las proteínas (tomada como la del PDBGGlobal). Curiosamente, ácido Aspártico, fuerte formador de cualquiera de los tres tipos de giros, muestra una propensión más baja a los giros Beta en las proteínas de *Thermus*, y en cambio propensiones más altas hacia la formación de dobleces y giros Beta, que la muestra PDBGGlobal y las de más proteínas agrupadas por organismo. La Asparragina (N) también presenta una menor tendencia a formar giros Beta que las demás proteínas. La Metionina en *Thermus*, presenta una propensión más alta hacia los giros al azar, que las propensiones que muestran las demás proteínas de esta agrupación.

Finalmente, en *Thermotoga*, otra bacteria termófila, la Histidina (H), presenta una propensión anormalmente alta hacia la formación de estructuras al azar, que en PDBGlobal y las demás proteínas de otros organismos. De hecho, es el aminoácido más fuerte formador de giros a azar, por encima incluso de la Prolina. Este aminoácido, Histidina, también presenta la menor propensión hacia la formación de estructuras helicoidales en proteínas de *Thermotoga*, aunque en las demás proteínas, es un mediano formador de hélices.

El Triptófano (W), presenta propensiones opuestas en *Thermus* y *Thermotoga*, hacia la formación de dobleces. En *Thermotoga* en cambio, el Triptófano tiende a tener una baja tendencia a los dobleces, mientras que en *Thermus*, tiene una propensión más alta. La Asparragina, por su parte, baja su tendencia a formar giros Beta, en las proteínas de *Thermus*.

Dado que *Thermus* y *Thermotoga* viven óptimamente en ambientes a temperaturas de 65°C y 80°C, respectivamente, lo que sugiere que la temperatura puede cambiar la tendencia intrínseca de algunos aminoácidos hacia la formación de una estructura secundaria particular (en este caso, C, D, H, N, P, W).

5.1.4. Propensiones de C&F en las proteínas agrupadas por Phylum.

En esta sección, se analizan las propensiones según el Phylum al que pertenecen los organismos de los que se aislaron y caracterizaron estructuralmente las proteínas. En este caso, y al igual que en las secciones anteriores de este capítulo, se analizaron las cinco estructuras secundarias, que se conocen como DSSP-5 (estructuras helicoidales, hilos/hojas Beta, dobleces, giros Beta y giros al azar).

En la figura siguiente (figura 5.1.4.1.), se consolidan las propensiones de las proteínas agrupadas por Phylum hacia las estructuras secundarias. En ella se puede notar que las propensiones presentan bastante homogeneidad, con muy pequeñas excepciones. Entre ellas, la Lisina (K), es el más fuerte formador de estructuras helicoidales en virus, mientras que en la generalidad de las proteínas, la muestra PDBGlobal, este aminoácido apenas se encuentra entre los medianos formadores, y los más fuertes formadores de hélices son Leucina, Glutámico y Alanina. La Glutamina (Q) y la Arginina (R), presentan también propensiones más fuertes hacia las hélices en las proteínas virales, que en los demás phyla.

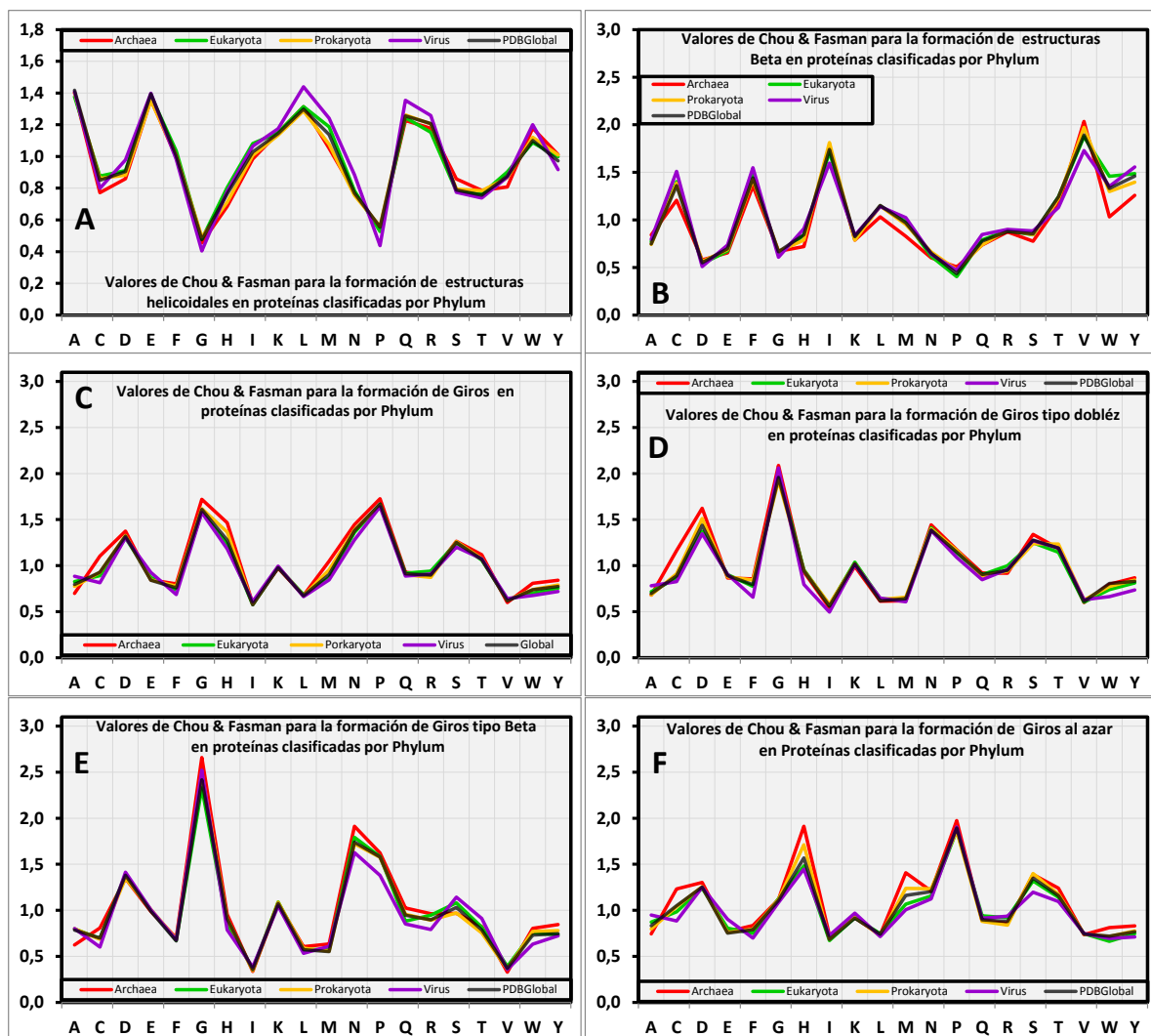


Figura 5.1.4.1. Propensiones Chou & Fasman de aminoácidos en proteínas agrupadas por el Phylum al que pertenecen.

En Archaea, el Triptófano (W), y la Tirosina (Y), presentan una menor tendencia a formar hilos y hojas Beta que sus propensiones en los demás phila. En el mismo grupo de organismos, Archaea, la Cisteína, la Histidina y la Metionina, presentan mayores propensiones hacia la formación de giros al azar, que lo que lo hacen en las proteínas de las demás agrupaciones de organismos.

5.1.5. Propensiones de C&F en las proteínas agrupadas por Tipo de enzima

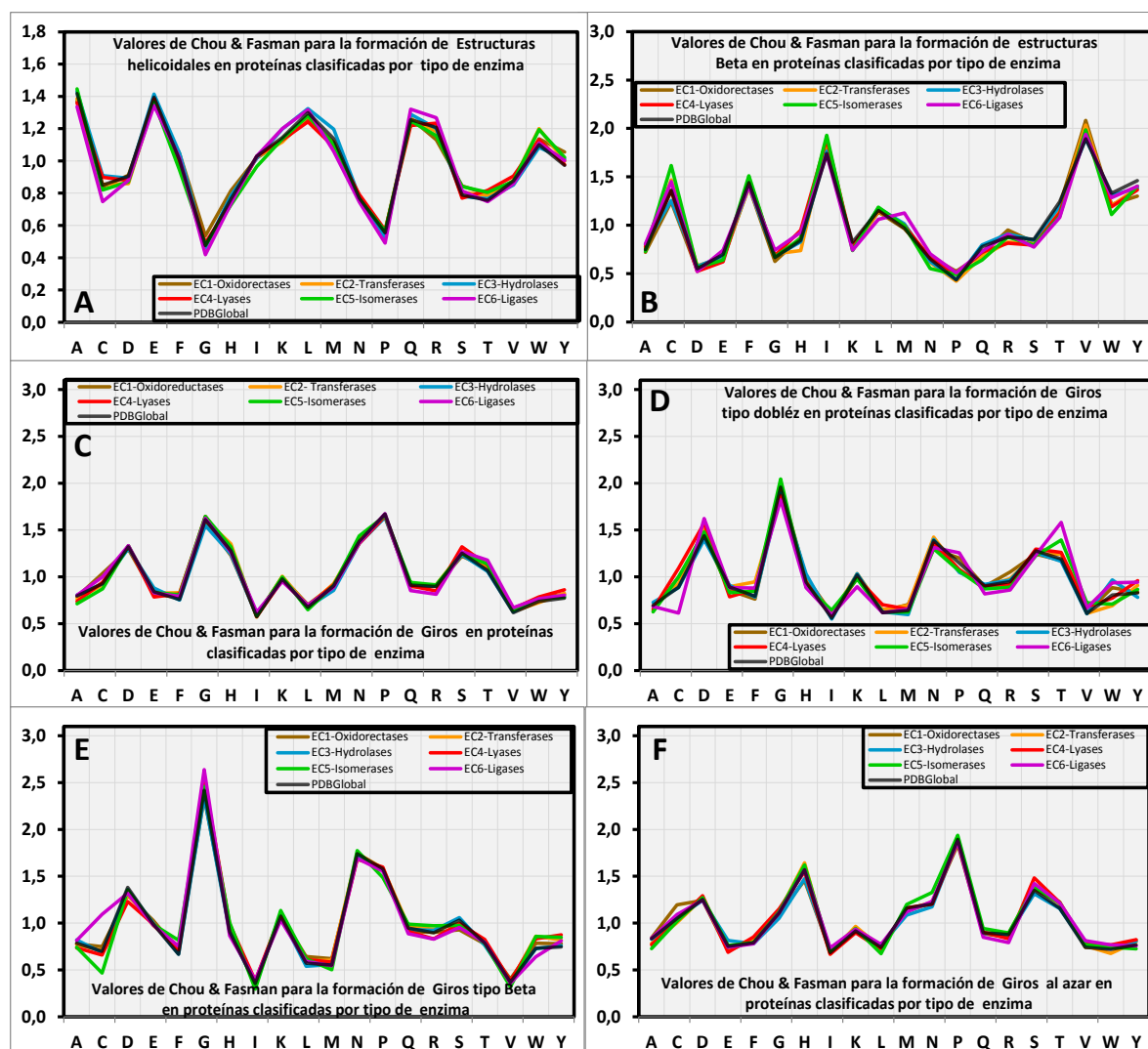


Figura 5.1.5.1. Propensiones Chou & Fasman de aminoácidos en proteínas agrupadas por el tipo de enzima, según la Comisión de Enzimas.

Como se puede ver en la figura 5.1.5.1., las agrupaciones de proteínas con actividad catalítica según el tipo de enzima, es decir, según el tipo de reacción que catalizan, muestran una mayor homogeneidad en las propensiones, que en los grupos vistos hasta el momento. Esto sugiere que la agrupación de proteínas en tipo de enzima, es una clasificación que deja de lado la estructura

secundaria, y en cierta medida, puede considerarse como “completamente artificial”, en la medida en que la agrupación no revela diferencias estructurales entre las proteínas.

Los dos grupos de enzimas que presentan las mayores diferencias en propensión, al igual que en las frecuencias, como se vió en el capítulo anterior, son las enzimas pertenecientes a los grupo 5 y 6 de la Comisión de Enzimas: las Isomerasas (EC-5) y las Ligasas (EC-6). Sin embargo, las diferencias en las propensiones, comparadas con la generalidad de las proteína (el PDBGlobal), son relativamente pequeñas, en comparación con otros grupos de proteínas.

A manera de comentario, el Triptófano (W), presenta menor tendencia a formar hilos Beta, que a formar estructuras helicoidales, que en las proteínas de los otros tipos de enzimas. Los residuos de Glutamina (Q) y de Arginina (R), en Isomerasas, presentan tendencias ligeramente mayores a formar hélices que cuando están en otras proteínas. Y la Cisteína en Isomerasas, muestra menor tendencia a formar giros Beta, mientras que en Ligasas una mayor tendencia a formar los mismos giros Beta, que en las proteínas Oxidorreductasas, Transferasas, Hidrolasas, Liasas y PDBGlobal.

5.1.6. Propensiones de C&F en las proteínas agrupadas por Familia proteica según PFam

Al igual que en el capítulo anterior, en que se compararon las frecuencias de aparición de residuos en estructura secundaria (sección 3.1.6), las proteínas agrupadas según la familia proteica a la que pertenecen, según los criterios de PFam, presentan la mayor heterogeneidad en sus valores de propensión hacia las diferentes estructuras secundarias.

Primero, es notorio que las propensiones de los 20 aminoácidos en las diferentes familias proteicas, hacia las cinco estructuras secundarias, son marcadamente diferentes.

Segundo, es claro que cuando se miran los giros totales, no se identifica tanta heterogeneidad como cuando se discriminan los diferentes tipos de giros, en que se observan diferencias claramente mayores entre las propensiones en diferentes familias.

Tercero, las diferencias entre familias proteicas son tan marcadas, que en este caso fue necesario cambiar la escala del eje y, que corresponden a la magnitud de la propensión. La escala, en las gráficas de las agrupaciones proteicas anteriores, secciones 4.1.1. a 4.1.5., están a una escala de propensión de 3.0, mientras que en las proteínas por familia proteica, están al doble, 6.0, como propensión.

En cuanto a las estructuras helicoidales, varios residuos presentan propensiones más bajas hacia estas estructuras, en algunos casos, disminuyendo hasta en más de un 90% la tendencia a formar estructuras helicoidales. Son estos los casos de Cisteína, Asparragina y Arginina en GFP, una conocida proteína todo Beta; Alanina, Glicina, Leucina, Prolina y Serina en proteínas del prión.

Por otro lado, salta a la vista, que el ácido Glutámico en las proteínas Anticongelantes (AFP), presenta una alta tendencia a formar estructuras helicoidales. Con una propensión aún mayor que el anterior residuo, el Triptófano en Proteína Verde Fluorescente (GFP), aparece como el más fuerte formador de estructuras helicoidales, con una propensión de más de 3.5 hacia las hélices. Esto lleva de nuevo a considerar que el concepto de propensión de Chou y Fasman, es un concepto dinámico y dependiente de las condiciones en que se encuentre la proteína, y el entorno de secuencia en que se encuentran el aminoácido.

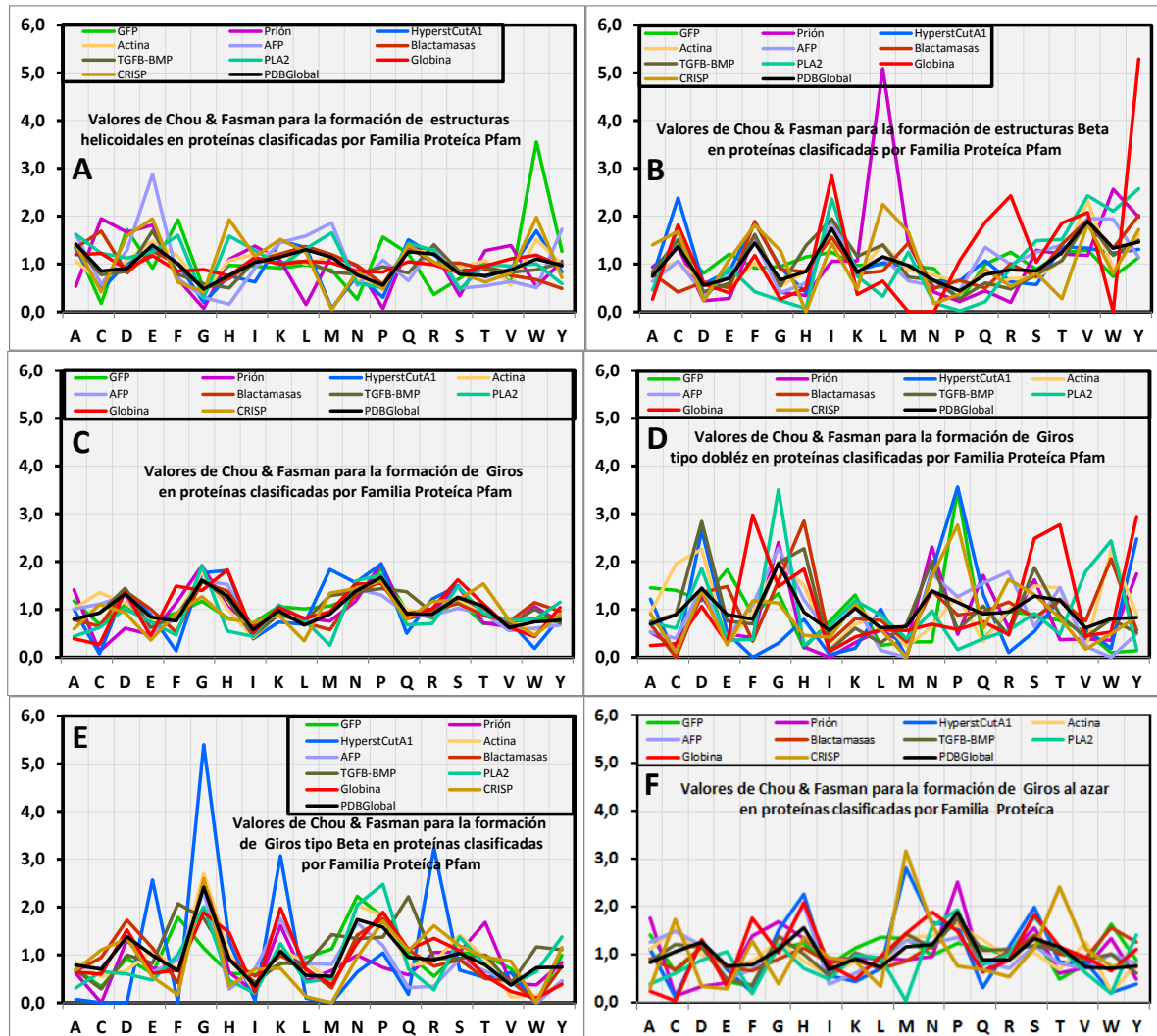


Figura 5.1.6.1. Propensiones Chou & Fasman de aminoácidos en proteínas agrupadas por familia proteica según PFam.

Un caso similar ocurre en las tendencias a la formación de hilos y hojas Beta. En las proteínas de Prión, el residuo de Leucina presenta una propensión de más de cinco veces más alta que su propensión hacia hojas en las demás proteínas del estudio. De nuevo, esto sugiere que el concepto clásico de propensión, debe ajustarse teniendo en consideración factores como el

entorno de secuencia del aminoácido y eventualmente otras condiciones particulares de la proteína.

Siguiendo con los comentarios sobre las propensiones en familias proteicas, la Histidina, tanto en la familia de proteínas CRISP, como en la de PLA2, presenta valor de propensión cero hacia la formación de hilos/hojas Beta.

5.2. Comparaciones entre Propensiones de Chou y Fasman, Entropía Conformacional y el Indicador de Preferencia Estructural

La Entropía estructural, es un valor que se ha calculado usando la ecuación de Shannon (Strait y Dewey, 1996). La formulación de la entropía de información está dada por:

$$S = -\sum_{i=1}^n p_i (\ln p_i)$$

Donde i se refiere en nuestro caso a uno de los tipos de estructuras secundarias analizados (hélices, hojas, giros Beta, dobleces, random coils), p_i es la probabilidad de encontrar la estructura secundaria i en el conjunto de proteínas; n , el número total de posibles estructuras secundarias (20; S es la entropía conformacional de cada aminoácido. Para facilitar la visualización, se normalizó la Entropía estructural para que sus valores encajen entre 0 y 1 (dividiendo por su valor máximo alcanzado), al igual que las propensiones de Chou y Fasman.

La Entropía conformacional pretende indicar en qué medida un aminoácido ó una secuencia de aminoácidos, tiende a ser la única formadora de una estructura secundaria (caso de baja entropía), ó por el contrario, tiende a formar diferentes estructuras secundarias (entropía alta).

El Indicador de Preferencia estructural, en adelante IPE, presentado por nosotros anteriormente (Mejía y Tischer, 2007), se calculó con base en la mayor frecuencia de aparición de un aminoácido multiplicado por la diferencia entre la mayor frecuencia y la segunda mayor frecuencia). Con esto, si un aminoácido aparece la totalidad de las veces en una estructura secundaria, su frecuencia en esta estructura sería de 1.0, y la segunda mayor frecuencia sería 0.0. En este caso, el IPE sería de 1.0. En cambio, si aparece con la misma frecuencia al menos en dos estructuras secundarias, el IPE sería de 0.0, indicando que no hay una preferencia hacia una estructura secundaria por parte del aminoácido, la tripleta, ó en general, la secuencia de aminoácidos. El IPE se calcula entonces así:

$$\text{IPE} = F_{\text{ma}} * (F_{\text{ma}} - F_{\text{sma}})$$

F_{ma} : Frecuencia más alta. F_{sma} : Segunda frecuencia más alta. IPE: Indicador de Preferencia estructural.

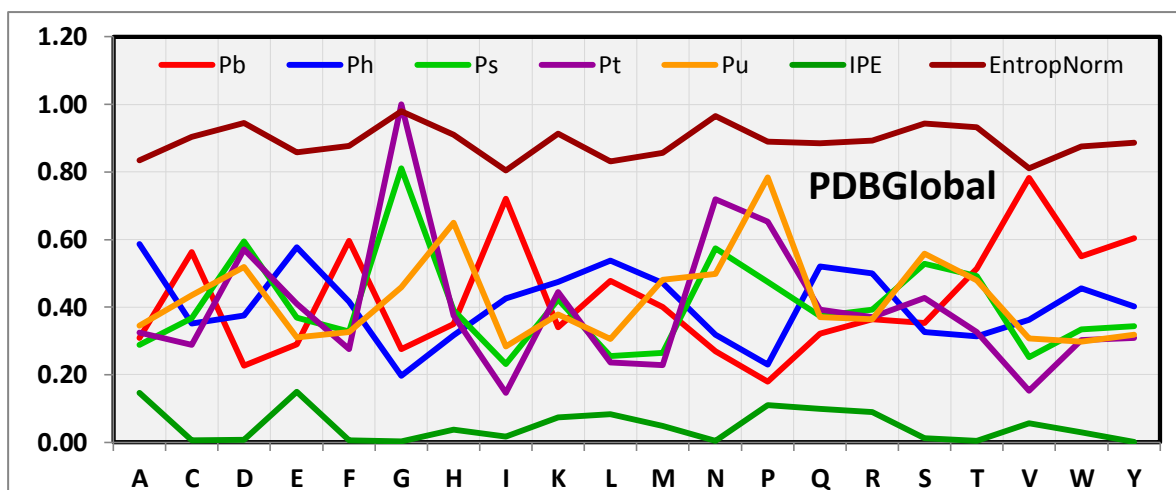


Figura 5.2.1. IPE, Entropías y Propensiones en los aminoácidos de PDBGlobal (para ajustar al rango entre cero y uno, se normalizaron las entropías y las propensiones).

Tendencias generales y particularidades

Propensiones hacia la formación de estructuras helicoidales (línea azul)

Aunque el valor de propensión no es especialmente alto, los mayores formadores de estructuras helicoidales son **Glutámico** y **Alanina** cuyas propensiones hacia otras estructuras son mucho menores. También, y en menor medida, son formadores de hélices, la **Leucina**, **Asparragina** y **Arginina**. Los residuos menos formadores de hélices, son la **Glicina** y la **Prolina**, fuertes formadores de giros.

Propensiones hacia la formación de hilos y hojas Beta (línea roja)

El mayor formador de hilos y hojas Beta, es la **Valina**, seguida, curiosamente, por la **Isoleucina** (curiosamente, porque un isómero de ella, la Leucina, es alto formador de hélices), y la **Tirosina**. Ambos aminoácidos presentan valores bajos de propensión hacia cualquiera de las otras estructuras secundarias. Los residuos menos formadores de hilos y hojas Beta, son la **Prolina** (fuerte formador de giros al azar), el **Glutámico** (formador de hélices) y la **Glicina** (formador de giros Beta y dobleces).

Propensiones hacia la formación de dobleces (línea verde claro)

Además de la **Glicina**, como fuerte formador de este tipo de estructura secundaria, y en menor medida, están el **Aspártico** y la **Asparragina**. Sin embargo, estos dos últimos residuos también presentan una propensiones similarmente altas, hacia otras estructuras secundarias, con lo que consideramos que necesitan la cercanía de otros residuos fuertemente formadores de esta estructura, para formarla.

Propensiones hacia la formación de giros Beta (línea morada)

Además de la **Glicina**, las **Asparragina** y en menor medida, la **Prolina**, pueden formar los giros Beta. Los menores formadores de estos giros, son la **Isoleucina** (formador de hojas Beta), y la

Valina (formador de hojas Beta). También son pobres formadores de giros Beta, la **Leucina** y la **Metioina**.

Propensiones hacia la formación de giros al azar (línea amarilla)

Los residuos más formadores de esta estructura, son la **Prolina** y en menor medida, la **Histidina**. Aparece también la **Serina** como formador de esta estructura, aunque también puede formar, casi en la misma medida, dobleces.

Los que no presentan altas propensiones hacia ninguna estructura secundaria

Se han identificado como fuertes formadores de estructuras secundarias, a la Alanina, el Glutámico, la Leucina, Glutamina y Arginina (hélices), Valina, Isoleucina, Fenilalanina y Tirosina (hilos y hojas Beta), Glicina (dobleses, giros Beta), Prolina, Histidina y Serina (Giros al azar). Sin embargo, hay unos pocos residuos de aminoácidos que no presentan una fuerte propensión hacia ninguna estructura y podríamos considerarlos una especie de residuos “comodín”, en el sentido de que se “acomodan” fácilmente en cualquier estructura secundaria, por lo que seguramente son más dependientes del contexto de secuencia de sus aminoácidos vecinos, y son ellos: **Cisteína** y **Triptófano** (que en el capítulo anterior mostraban bajas e iguales frecuencias de aparición en cualquier estructura secundaria), **Glutámico**, **Lisina**, **Metionina**, **Glutamina**, y **Treonina**. Como se verá en la sección de análisis de propensiones, entropías e IPE para tripletas, estos residuos “comodín”, tienden a formar lo que hemos llamado “Tripletas camaleón”, es decir, tripletas de aminoácidos que se acomodan igualmente en cualquier estructura secundaria, sin presentar preferencia dominante hacia ninguna de ellas.

En cuanto a la Entropía (línea marrón) y el IPE (línea verde oscuro)

Llama la atención que el valor de la entropía, parece comportarse como un inverso aritmético del IPE. Sin embargo, este comportamiento inverso no se conservó al comparar las entropías y el IPE en tripletas, con lo que nuestro IPE se diferencia de la entropía conformacional. Por otro lado, una ventaja que ofrece nuestro indicador, es que en el cálculo, lo podemos asociar con la estructura secundaria que más frecuentemente forma un aminoácido ó una tripleta, lo que no ocurre con la entropía.

En cuanto al valor de IPE, los residuos con valores más altos de IPE, resultaron ser Alanina y Glutámico, fuertes formadores de estructuras helicoidales. También la Prolina. Este indicador está hablando no sólo sobre propensión hacia una estructura secundaria, sino también de la propensión hacia una única estructura secundaria. Esto significa que si un aminoácido es fuerte formador de dos estructuras secundarias, su valor de IPE es bajo. A manera de ejemplo, la Glicina, que hipotéticamente debería tener una alta preferencia estructural, al discriminar en diferentes tipos de giros, su valor de IPE se baja considerablemente. De aquí, que consideramos el valor de IPE, como un indicador de la “fortaleza” con que un aminoácido ó una tripleta forma una estructura secundaria y por tanto puede ser usado para identificar zonas más estables ó menos estables hacia la formación de estructuras secundarias en una proteína. Este concepto se amplía y se hace más marcadamente útil en el estudio de tripletas de la sección siguiente.

Propensiones, Entropía e IPE en proteínas agrupadas por su contenido de estructura

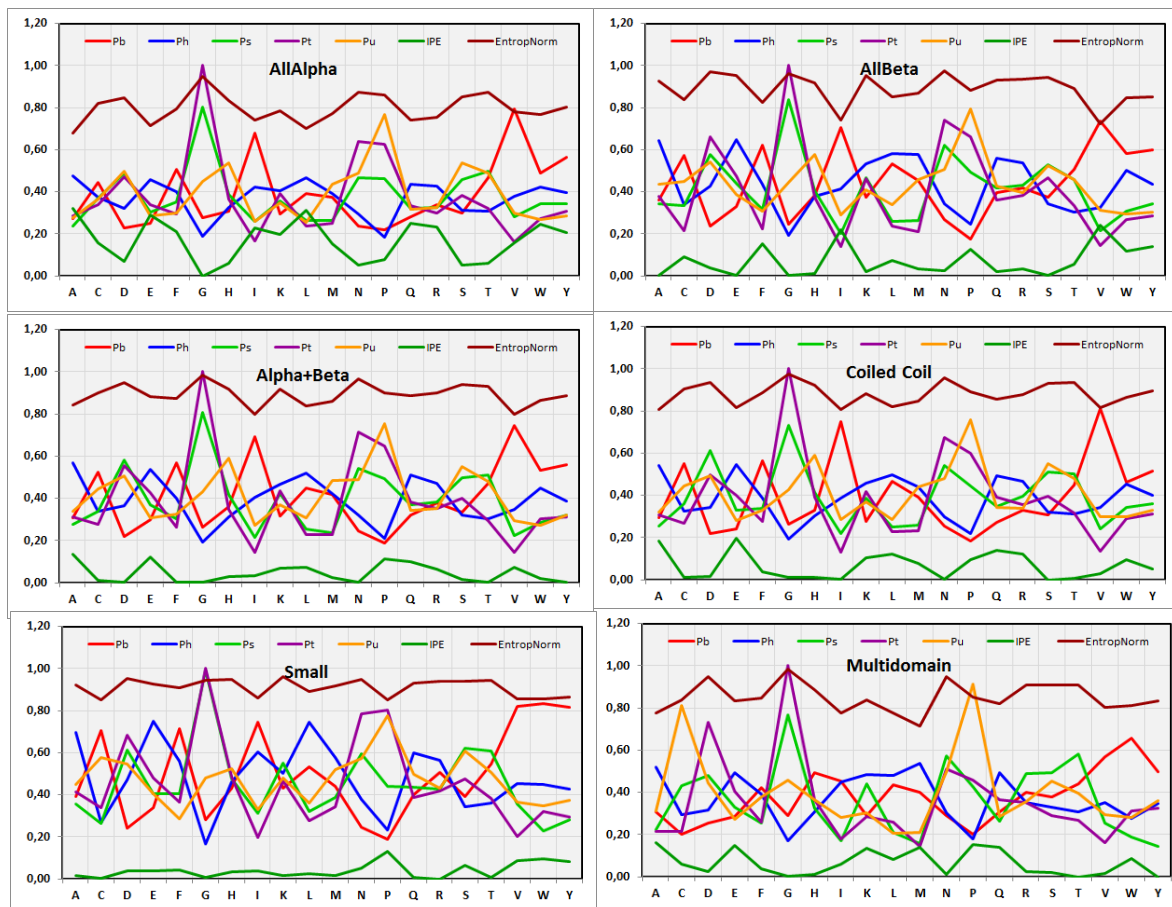


Figura 5.2.2. IPE, Entropías y Propensiones en los aminoácidos en proteínas agrupadas por contenido de estructura secundaria según SCOP.

Tendencias generales y particularidades

Entre las generalidades más conspicuas, están: La Glicina es el residuo que más fuertemente tiende a formar giros Beta; si embargo, en algunos grupos de proteínas, sus valores hacia la formación de dobleces y de los valores de Propensión, Entropías e IPE no cambian mucho con respecto al PDBGlobal. Hay diferencias particulares, como que la Valina (V), el Triptófano (W) y la Tirosina (Y), presentan las más altas propensiones hacia la formación de hilos y hojas Beta en las proteínas clasificadas como “Small”. Aunque en general, la Prolina (P) es el mayor formador de giros al azar, en el caso de las proteínas multidominio, este residuo es aún más fuerte formador de estas estructuras. En esas proteínas, también aparece la Cisteína (C) como un muy fuerte formador de giros al azar, que en los demás grupos de proteínas de SCOP, no presenta una

propensión tan alta. En todos los grupos de proteínas, la mayor propensión la presenta la Glicina, que en todas las proteínas agrupadas según SCOP, alcanza la máxima propensión hacia la formación de giros Beta. El mismo residuo de aminoácidos, presente bajas propensiones hacia las estructuras helicoidales y los hilos Beta. A manera de contraste, la Histidina (H) y la Lisina (K), presentan bajas propensiones hacia cualquier estructura secundaria, lo que sugiere que estos residuos de aminoácido funcionan como una especie de residuo “comodín”, que se acomoda en cualquier estructura secundaria, lo que sugiere que dependen del contexto de estructura secundaria que forman los demás residuos en una proteína.

Las proteínas Todo-Alfa son el grupo de proteínas que presenta los mayores valores de IPE, y los menores valores de Entropía, lo que significa que son las proteínas en que los residuos de aminoácidos tienden a aparecer más marcadamente formando estructuras secundarias específicas. En diferentes grupos de proteínas de SCOP, los aminoácidos presentan diferentes valores de IPE, es decir, que estos aminoácidos con valores altos de IPE, aparecen predominantemente en una estructura secundaria, lo que sugiere que presentan una alta tendencia a formar una estructura secundaria en particular. El valor menor de IPE, se encuentra en las proteínas “Small”, en que ningún aminoácido parece presentar una frecuencia predominantemente alta en ninguna estructura secundaria, con la excepción de la Prolina (P), que aparece más en giros al azar que en otras estructuras secundarias.

El mayor valor de IPE en todos los grupos de proteínas según SCOP, lo tiene la Prolina, que en las proteínas “Todo Beta”, aparece formando más frecuentemente hilos y hojas Beta, que ninguna otra estructura secundaria.

5.3. Propensiones de Chou y Fasman, Entropías e IPE en tripletas de aminoácidos

En esta sección, se analizan las propensiones de Chou y Fasman, y se compara la Entropía Conformacional, con el Indicador de Preferencia Estructural para tripletas de aminoácidos.

Como uno de los aportes del análisis de tripletas, es que permite no sólo identificar los residuos con alta propensión hacia una estructura secundaria, sino que también permite hacerlo para las transiciones entre estructuras secundarias, por lo que en las siguientes figuras, hemos optado por presentar las propensiones calculadas según Chou y Fasman, en términos de figuras de tripletas estructurales puras y transicionales. Las propensiones de Chou y Fasman, se normalizaron para que sus valores encajaran entre 0 y 1, y esto permitiera hacer comparaciones con los valores de Entropía (también normalizada), e IPE. Para facilitar la visualización, en las figuras siguientes se presentan sólo las barras que superan el valor de propensión de 0.6 hacia cada estructura secundaria. Escogimos un umbral alto, para facilitar la interpretación, la identificación de tripletas

con mayor propensión, y disminuir el ruido que producen tripletas con menores valores de propensión.

Generalizando las definiciones a tripletas de aminoácidos XYZ y tripletas estructurales ABC se define:

Propensión:

$$\frac{(\# \text{ de veces que aparece XYZ en ABC} / \text{frecuencia total de ABC})}{(\# \text{ total de apariciones de XYZ} / \# \text{ total de tripletas de aminoácidos})}$$

Entropía Conformacional S:

$$S = -\sum_{i=1}^n p_i (\ln p_i)$$

donde i se refiere a las diferentes tripletas estructurales y p_i es la probabilidad de encontrar la tripleta estructural i en el conjunto de proteínas.

Indicador de preferencia estructural IPE para una tripleta de aminoácidos XYZ:

$$\text{IPE} = F_{\text{ma}} * (F_{\text{ma}} - F_{\text{sma}})$$

F_{ma} : Frecuencia más alta entre las tripletas estructurales adoptadas por la tripleta de aminoácidos XYZ;

F_{sma} : Segunda frecuencia más alta adoptados por la tripleta XYZ.

Aquí calculamos las propensiones en las combinaciones de tres residuos de aminoácidos, lo cual nos llevó a establecer las estructuras secundarias, y combinaciones de las mismas, para tres residuos de aminoácidos. A estas tripletas de estructura secundaria, las hemos llamado “tripletras estructurales”, e incluyen tripletas puras, que están en una sola estructura secundaria, de la forma HHH para una tripleta en estructura helicoidal; tripletas transicionales, que corresponden a zonas de transición entre diferentes estructuras secundarias, como transiciones entre estructuras helicoidales y giros beta, por ejemplo, de la forma HTT ó HHT; y finalmente, tripletas de difícil interpretación biológica, de la forma HTB.

5.3.1. Las Propensiones en tripletas de aminoácidos

Al pasar de aminoácidos individuales a tripletas, aparecen en las tripletas estructurales correspondientes, aparte de las tripletas puras (como HHH o EEE por ejemplo), tripletas que combinan estructuras secundarias (como HEE o HEH). Entre ellas aparecen las tripletas que llamamos transiciones entre estructuras (como HHE o HEE). En lo que sigue se analizan para cada dos estructuras las correspondientes tripletas puras junto con las transiciones. Las tripletas mixtas que no son puras o transicionales (como HEH o HEU) se excluyen del presente análisis dado que son difíciles de interpretar.

Hélices, dobleces y transiciones entre estas estructuras

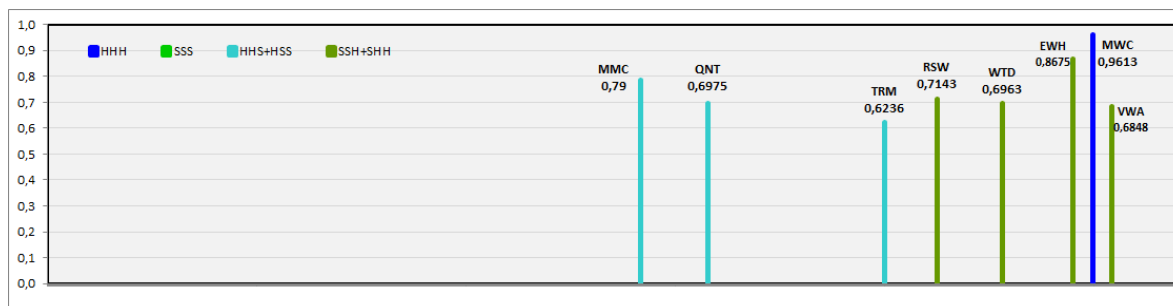


Figura 5.3.1. Propensiones hacia la formación de hélices (HHH), dobleces (SSS), y las transiciones de hélice a dobléz (HHS+HSS), y de dobléz a hélice (SSH+SHH).

En la anterior figura, se puede ver que la tripleta más formadora de hélices es MWC (Metionina-Triptófano -Cisteína), formada por residuos que no son conocidos por tener alta propensión hacia la formación de hélices. Este es un hallazgo curioso e inesperado para nosotros, dado que los más fuertes formadores de hélices deberían estar formando las tripletas con mayores propensiones hacia esta estructura. Esta combinación de residuos que hemos llamado “comodín”, por su baja propensión hacia algunas estructuras secundarias, puede presentar una alta propensión hacia hélice, por encontrarse posiblemente en medio de residuos fuertes formadores de hélices, con lo que el contexto los ha llevado a formar la misma estructura de su contexto, una hélice.

En la figura no aparecen tripletas formadores de dobleces (SSS), ya que todas estas tienen una propensión encima de 0.6, por lo que se suprimen en la gráfica.

En cuanto a la transición de hélices a dobléz (HHS+HSS), aparecen las tripletas MMC (Metionina-Metionina-Cisteína), QNT (Glutamina-Asparagina-Treonina), TRM (Treonina-Asparagina-Metionina), RSW (Arginina-Serina-Triptófano) y WTD (Triptófano-Treonina-Aspártico). Asparagina (N), Serina (S) y Aspártico (D), ya habían mostrado altas propensiones hacia los dobleces, con lo que era más ó menos de esperar que aparecieran en las tripletas de transición de hélices a dobléz. Sin embargo, los residuos de Metionina, Cisteína, Glutamina, Treonina y Triptófano, no mostraban alta propensión hacia esta estructura; y estos tienen una tendencia alta hacia la aparición en estas transiciones. Esto nos lleva a pensar que la propensión de residuos individuales de aminoácidos, está perdiendo parte del panorama, y que los residuos con bajas propensiones hacia las estructuras “puras” (de la forma HHH, BBB, SSS, TTT y UUU), pueden ser residuos con propensión a hacer las transiciones entre estructuras secundarias. En este sentido ofrecemos con el estudio de tripletas transicionales un enfoque que generaliza el análisis de las secuencias N-cap y C-cap que comienzan y terminan hélices (Doig y Baldwin, 1995; Kruus et al, 2005).

Hélices, dobleces y transiciones entre estas estructuras

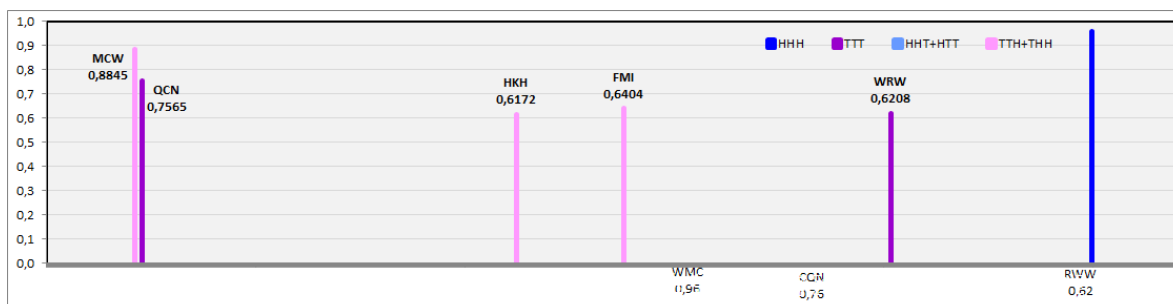


Figura 5.3.2. Propensiones hacia la formación de hélices (HHH), giros Beta (TTT), y las transiciones de hélice a giro Beta (HHT+HTT), y de giro Beta a hélice (TTH+THH).

Las propensiones más altas hacia la formación de dobléz, estuvieron en las tripletas QCN (Glutamina-Cisteína-Arginina), y WRW (Triptófano-Arginina-Triptófano), mientras que la transición entre Giros Beta y hélices, están principalmente formadas por MCW (Metionina-Cisteína-Triptófano), FMI (Fenilalanina-Metionina-Isoleucina), y HKH (Histidina-Lisina-Histidina).

Hélices, giros al azar, y transiciones entre estas estructuras

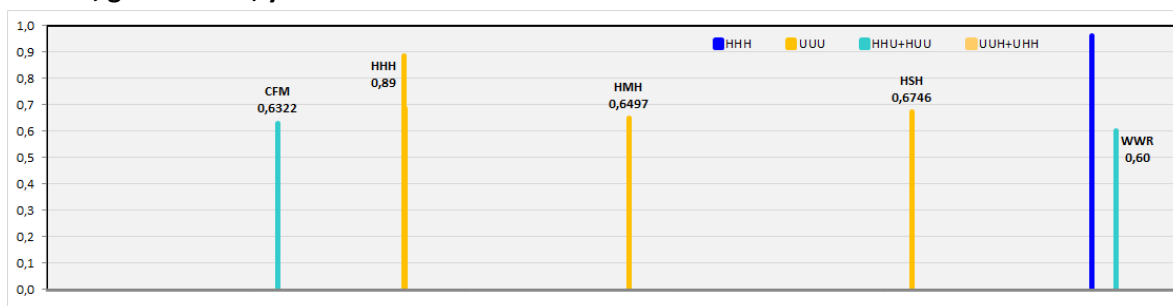


Figura 5.3.3. Propensiones hacia la formación de hélices (HHH), giros al azar (UUU), y las transiciones de hélice a dobléz (HHU+HUU), y de dobléz a hélice (UUH+UHH).

En la figura anterior, se puede ver que la tripleta que más alta propensión presenta hacia los giros al azar, son las tripletas de Histidinas (HHH), seguidas por combinaciones del mismo residuo, como HHH (Histidina-Metionina-Histidina) y HSH (Histidina-Serina-Histidina). HHH es la tripleta que encontramos como más frecuente en el PDB, y según la propensión, la mayoría de las veces, aparece en estructuras al azar, con lo que su valor de IPE será también alto. Esto significa que esta tripleta, HHH, es una de las combinaciones de residuos de aminoácido, que con mayor certeza puede uno usar para la construcción de giros aleatorios en proteínas.

La transición entre hélices y giros al azar, está dada principalmente por la combinación de residuos CFM (Cisteína-Fenilalanina-Metionina) y WWR (Triptófano-Triptófano-Arginina).

No aparecen propensiones mayores a 0.6 (normalizadas), hacia las transiciones entre giros al azar y hélices.

Hélices, hilos/hojas, y transiciones entre estas estructuras

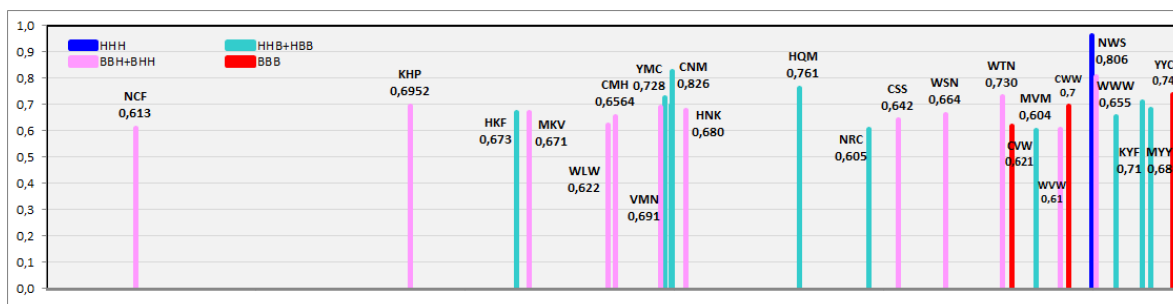


Figura 5.3.4. Propensiones hacia la formación de hélices (HHH), hilos y hojas Beta (BBB), y las transiciones de hélice a hilos (HHB+HBB), y de hilos a hélice (BBH+BHH).

Es llamativo que en la transición de hilos/hojas Beta hacia hélices, hay un gran número de triptetas con propensiones por arriba de 0.6. Las triptetas que forman estas transiciones, son NCF, KHP, MKV, WLW, CMH, VMN, HNK, CSS, WSN, WTN, WVV, y NWS. Es decir, triptetas ricas en Triptófanos, y combinaciones con otros aminoácidos como Valina, Treonina, Asparragina, y Lisina. Otras triptetas contienen Cisteína, Histidina y Prolina.

Hay 9 triptetas que forman fuertemente la transición entre hélices e hilos/hojas Beta: HKF, YMC, CNM, HQM, NRC, CVW, KYF, MYV y WWW.

Hilos/Hojas, dobles, y transiciones entre estas estructuras

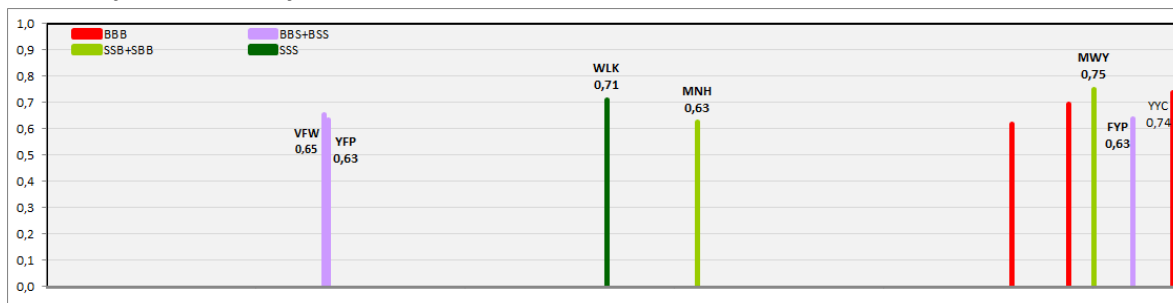


Figura 5.3.5. Propensiones hacia la formación de hilos y hojas Beta (BBB), dobles (SSS), y las transiciones de hilos/hojas a dobles (BBS+BSS), y de dobles a hilos (SSB+SBB).

Como ya se han comentado las propensiones hacia la formación de hilos/hojas y de dobles, se comenta entonces sobre las transiciones de hoja a dobléz y viceversa. En la transición de hoja a dobléz, las triptetas con mayor propensión a aparecer en esta zona de la estructura secundaria. En cuanto a esto, tres triptetas marcan la transición entre hilos/hojas y dobles: VFW, YFP y FYP, estando la Fenilalanina como constante y eventualmente la Prolina participando de la transición. Es ampliamente conocido que la Prolina tiende a terminar hojas, y que no puede ser un donador de enlace de hidrogeno, dado que no lo tiene en su grupo amida. Así que las Prolinas pueden interrumpir hojas, preferiblemente hacia dobles, ó hacia giros al azar.

En la transición de dobles a hilos/hojas, encontramos las triptetas con las mayores propensiones hacia esta estructura: MNH y MWY. La Metionina (M), en forma de oligómero, puede formar hilos

Beta con torsiones, lo que puede sugerir que por ello aparezca en estas transiciones (Paskowskia et al, 1978).

Hilos/Hojas, giros Beta, y transiciones entre estas estructuras

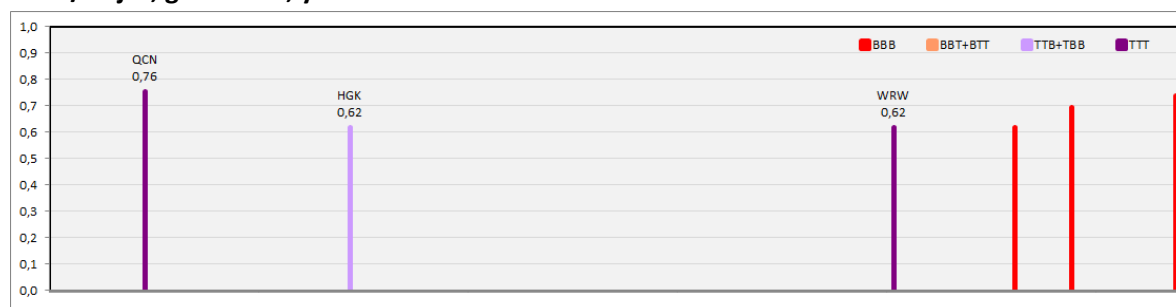


Figura 5.3.6. Propensiones hacia la formación de hilos y hojas Beta (BBB), giros Beta (TTT), y las transiciones de hilos/hojas a dobleces (BBT+BTB), y de giros Beta a hilos (TTB+TBB).

Las transiciones entre hilos/hojas y giros Beta, no presentan una tripleta con propensión mayor a 0.6. Las transiciones entre Giros Beta e hilos/hojas, está dada por la tripleta HGK, con la mayor propensión a esta estructura.

Hilos/Hojas, giros al azar, y transiciones entre estas estructuras

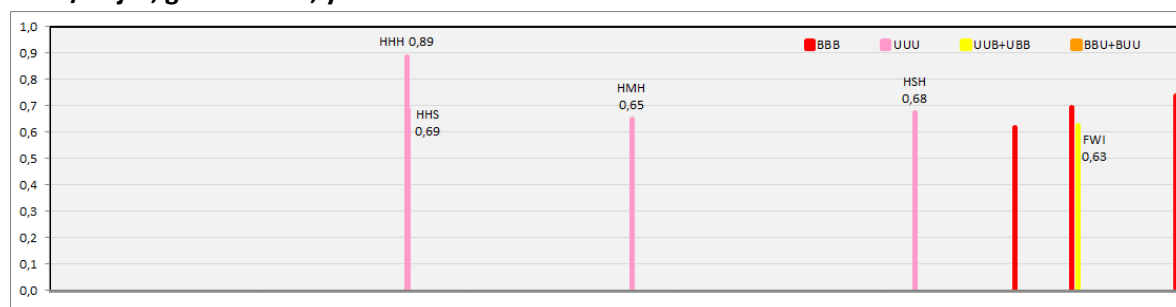


Figura 5.3.7. Propensiones hacia la formación de hilos y hojas Beta (BBB), giros al azar (UUU), y las transiciones de hilos/hojas a giros aleatorios (BBU+BUU), y de giros Beta a hilos (UUB+UBB).

Las transiciones de hilos/hojas a giros al azar, no tiene una tripleta con propensión suficientemente alta (al menos de 0.6). Solo la tripleta FWI muestra una propensión alta hacia la transición entre giros al azar y los hilos Beta.

Dobléz, giros Beta y sus transiciones

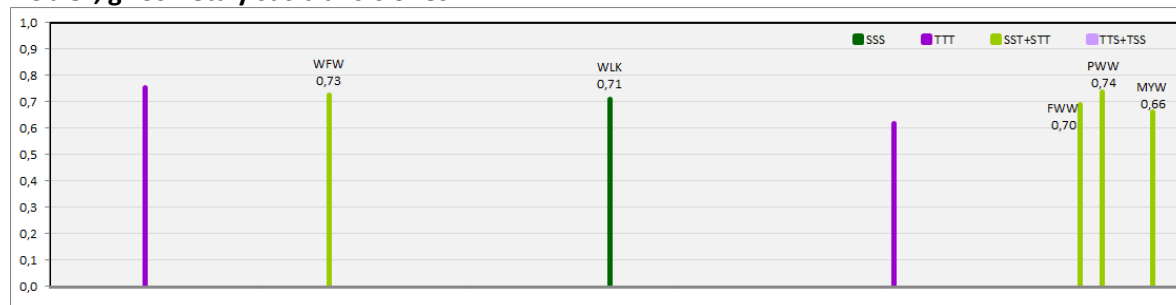


Figura 5.3.8. Propensiones hacia la formación de dobleces (SSS), giros Beta (TTT), y las transiciones de dobleces a giros Beta (SST+STT), y de giros Beta a dobléz (TTS+TSS).

Las transiciones de giros Beta a dobleces, no tienen una tripleta con propensión suficientemente alta (al menos de 0.6). Solo las tripletas WFW, FWW, PWW y MYW muestran una propensión alta hacia la transición entre dobleces y los giros Beta.

Giros Beta, giros al azar, y transiciones entre estas estructuras

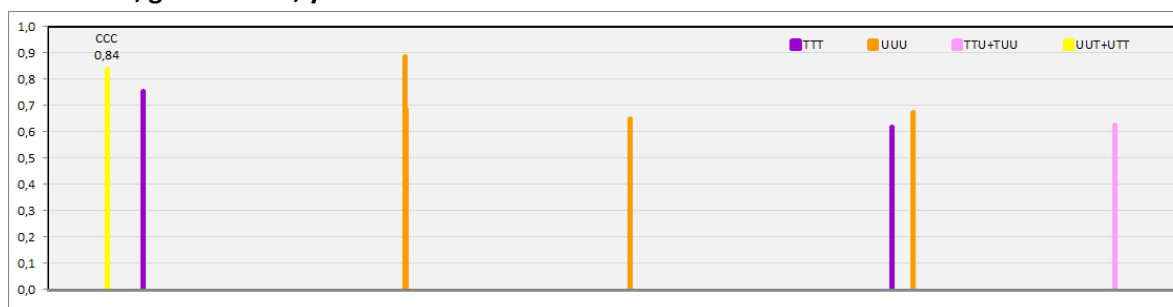


Figura 5.3.8. Propensiones hacia la formación de giros Beta (TTT), giros al azar (UUU), y las transiciones de hilos/hojas a giros aleatorios (TTU+UUU), y de giros Beta a hilos (UUT+UTT).

La transición entre giros al azar y giros Beta, está dada por una tripleta formada por Cisteínas (CCC). Desde una perspectiva evolutiva y de funcionalidad, esta tripleta tiene sentido como giro que aflora a la superficie de una estructura proteica y le permite establecer enlaces de disulfuro intramonómero o intermonómeros.

5.3.2. Las Entropías y el Indicador de Preferencia Estructural en tripletas de aminoácidos

Es importante dejar en claro que el valor de IPE no habla de la propensión de una tripleta ó aminoácido hacia la formación de una estructura secundaria, sino del grado de “absolutez”, con que una tripleta forma estructuras secundarias. A manera de ejemplo usaremos como ejemplo un residuo que presenta propensiones altas hacia la formación de giros tipo dobléz, y de giros tipo Beta, como la Glicina. Sin embargo, el hecho de que aparezca en dos estructuras secundarias con alta frecuencia (pues la propensión es en últimas, una normalización de la frecuencia de aparición en una estructura), hará que el valor de IPE para la Glicina, sea bastante bajo. En cambio, un aminoácido o una tripleta que aparezca sólo en una estructura secundaria, tendrá un alto valor para IPE.

Por otro lado, cuando analizamos las propensiones e IPE en los grupos de proteínas encontramos que su valor como indicador se pierde un poco, dado que en algunos grupos de proteínas, muchas tripletas no se han usado para construir las proteínas del grupo, por lo que no habría valor de IPE para ellas. Debido a esto, optamos por analizar y usar el IPE globalmente.

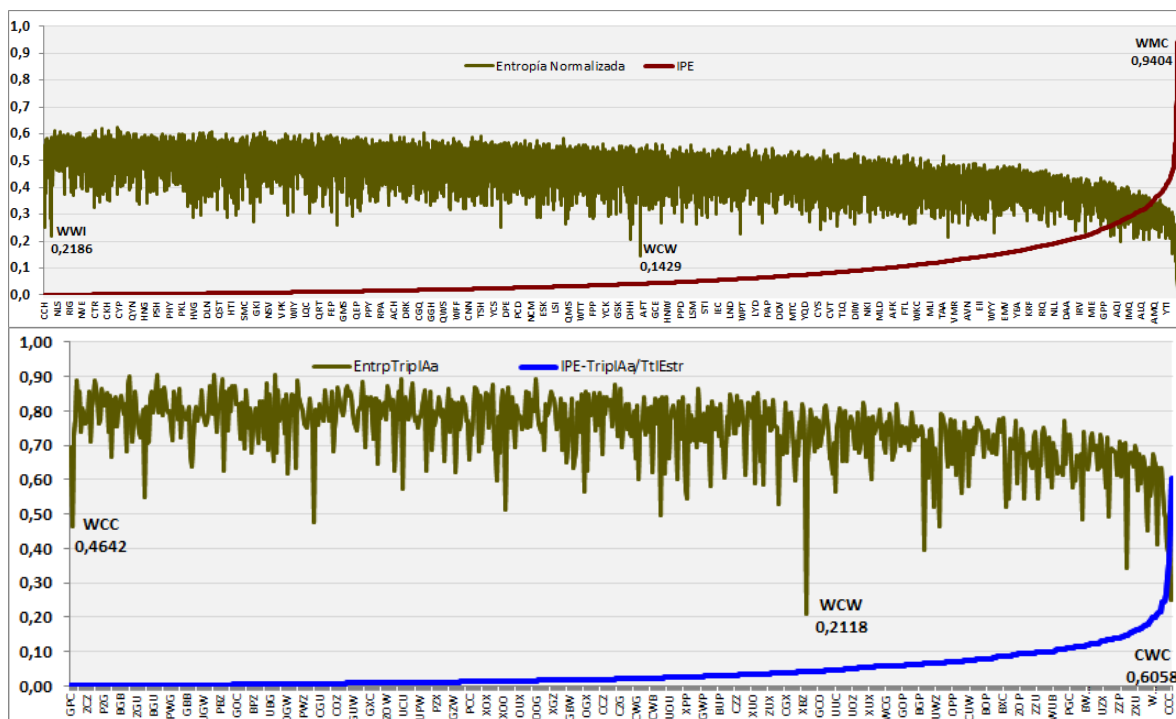


Figura 5.3.2.1. Comparación entre el IPE calculado por tripletas de aminoácidos, y la Entropía Conformacional, también calculada por tripletas de aminoácidos.

En la figura anterior (5.3.2.1.), se presentan las comparaciones entre la Entropía Conformacional y el IPE calculados tanto para tripletas de aminoácidos, como para tripletas de alfabeto reducido. Aunque ambos tienen un comportamiento aproximadamente inverso, hay diferencias importantes, que nos llevan a considerar al IPE como nuestro valor a usar para la identificación de tripletas con mayor “fortaleza” en la tendencia a formar estructuras secundarias en particular. En las dos tablas siguientes,

Tabla 5.3.2.1. Tripletas con valores cero, de IPE, que considera tripletas camaleón.

Tripleta	IPE	Tripleta	IPE	Tripleta	IPE
CCH	0,0000	IPC	0,0000	PHA	0,0000
CCT	0,0000	IPD	0,0000	PWD	0,0000
CMC	0,0000	KCG	0,0000	QCG	0,0000
CNT	0,0000	KNR	0,0000	RGC	0,0000
CTC	0,0000	KWS	0,0000	THE	0,0000
CTW	0,0000	LPM	0,0000	TRN	0,0000
DHD	0,0000	MPM	0,0000	TWF	0,0000
DPC	0,0000	MRP	0,0000	VNC	0,0000
FEW	0,0000	MTD	0,0000	WCC	0,0000
FKG	0,0000	NDC	0,0000	WFW	0,0000
FWW	0,0000	NID	0,0000	WFI	0,0000
GPC	0,0000	NKD	0,0000	WHM	0,0000
GVN	0,0000	NMY	0,0000	WNW	0,0000
HGG	0,0000	NPC	0,0000	WTS	0,0000
HHC	0,0000	NSH	0,0000	WWI	0,0000
HYF	0,0000	PCM	0,0000	YIH	0,0000
HYH	0,0000	PCN	0,0000	YPR	0,0000

Tabla 5.3.2.2. Tripletas con valores altos de IPE, que consideramos son más determinantes de la estabilidad de la estructura secundaria.

Tripleta	IPE	Estructura	Tripleta	IPE	Estructura	Tripleta	IPE	Estructura
WMC	0,9404	HHH	VII	0,5414	BBB	VWV	0,4594	BBB
HHH	0,9116	UUU	NWM	0,5211	HHH	CWF	0,4582	BBB
SHH	0,7332	UUU	HSS	0,5209	UUU	EEA	0,4567	HHH
YYC	0,7270	BBB	HHM	0,5208	UUU	EMA	0,4542	HHH
MHH	0,7058	UUU	FQM	0,5095	HHH	SSH	0,4500	UUU
HHS	0,6239	UUU	EAM	0,5009	HHH	LRE	0,4477	HHH
CWC	0,6058	BBB	VCW	0,4921	BBB	VYY	0,4465	BBB
VVV	0,5865	BBB	YFC	0,4920	BBB	EAI	0,4463	HHH
VVI	0,5755	BBB	IVI	0,4913	BBB	MCL	0,4451	HHH
MWA	0,5724	HHH	EHH	0,4786	UUU	MGS	0,4445	UUU
VIV	0,5724	BBB	EAA	0,4721	HHH	QAL	0,4428	HHH
III	0,5693	BBB	VVC	0,4718	BBB	EQL	0,4397	HHH
IIV	0,5634	BBB	AQA	0,4708	HHH	EQI	0,4388	HHH
MWW	0,5586	BBB	EEL	0,4681	HHH	EML	0,4371	HHH
SHM	0,5565	UUU	RMC	0,4611	HHH	GHH	0,4365	UUU
RWY	0,5539	HHH	ELA	0,4601	HHH	VLV	0,4356	BBB
VYI	0,5449	BBB	EAL	0,4599	HHH	EEL	0,4320	HHH

En las dos tablas anteriores, se pueden ver las cincuenta tripletas que presentan un IPE de cero, que hemos llamado tripletas “camaleón”, pues su bajo IPE indica que no presentan preferencia hacia ninguna estructura secundaria en particular, mientras que también se presentan las cincuenta tripletas con los mayores valores de IPE.

Comparamos entonces dos proteínas, una conocida por su estabilidad en términos de la estructura: la Proteína Verde Fluorescente (GFP), y una proteína que cambia su plegamiento y se asocia con enfermedad causada por ese plegamiento incorrecto, el Prión.

La figura 5.3.2.2., muestra los valores de IPE de las tripletas que componen a las secuencias de las GFP (color verde), y los de las tripletas que componen las secuencias de las proteínas del Prión (color morado). En ella hemos encontrado que la proteína GFP está formada por tripletas con mayores valores de IPE, en comparación con los IPE de las tripletas que componen las proteínas del Prión comparado los valores de IPE de las tripletas que constituyen la secuencia de aminoácidos de las dos proteínas.

Los IPEs de las tripletas que componen a GFP, consistentemente más altos que los de las tripletas que se han usado para construir las proteínas del Prión, pueden contribuir a explicar la diferencia en estabilidad entre las dos familias de proteínas. Los IPE altos contribuyen a explicar que las GFP son proteínas más estables, con menor tendencia a cambiar su estructura, mientras que el Prión es una proteína que puede cambiar la estructura secundaria, y por tanto la estructura terciaria, y alcanzar estructuras tridimensionales que facilitan la agregación y la formación de cuerpos amiloides, causantes de enfermedad. El IPE estaría permitiendo identificar una baja estabilidad de la estructura secundaria de esta proteína.

Consideramos que el uso del IPE para la comparación de proteínas en términos de la identificación de su estabilidad estructural, es uno de los aspectos más interesantes de nuestro indicador.

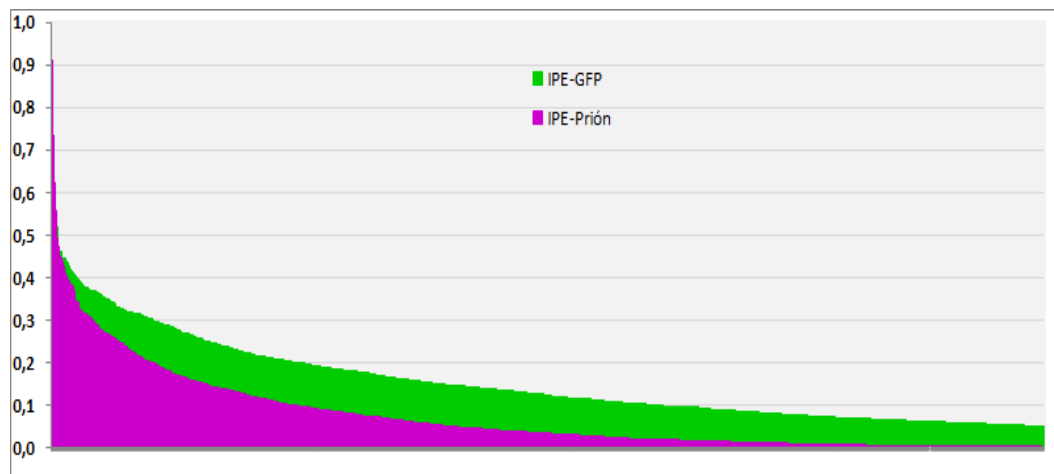


Figura 5.3.2.2. Comparación entre los IPE de las tripletas que componen a las Proteínas Verdes Fluorescentes y las proteínas del Prión. En el eje X, las tripletas que componen la proteína, y en el eje Y, los valores de IPE de las mismas.

Pero esta es una comparación entre dos proteínas con conocidos y diferentes niveles de estabilidad estructural. Por ello, hemos querido ir un poco más allá, ampliando las comparaciones con otras proteínas que se presenta a continuación.

Considerando que el máximo valor de IPE es de 1, y que en nuestra muestra PDBGlobal, el máximo valor de IPE hallado es de 94%, en la tripleta WMC, escogimos un umbral del 10% para considerar bajos IPE (es decir, IPE menor de 0.1), como indicadores de tripletas que aparecen en varias

estructuras secundarias sin una preferencia en particular, lo que hemos interpretado como tripletas que no determinan la estructura secundaria por sí mismas, y por tanto pueden cambiar su plegamiento dependiendo de las condiciones microambientales de la proteína, ó de la secuencia contexto en que están inmersas en la secuencia proteica.

En la siguiente tabla puede observarse que diferentes proteínas, están construídas con diferente composición de tripletas, y en últimas construídas por tripletas que muestran valores diferentes de IPE, lo que permite hacer inferencia sobre la estabilidad de las mismas.

En esa tabla, hemos considerado 1286 tripletas que constituyen a las proteínas presentadas aquí. Escogimos tripletas más estables, termoestables, como las de Thermus y Thermotogas, y otras proteínas.

La tabla está dividida en tres secciones de dos renglones cada una. En el renglón de arriba, se presenta el número de tripletas que presenta cada proteína, que cumple con la condición de la parte más a la izquierda (Umbral). El segundo renglón dice cuál es el porcentaje de esas tripletas con respecto al total de tripletas en el grupo de proteínas.

Tabla. 4.3.2.3. Umbrales de IPE y porcentaje de tripletas que constituyen diferentes tipos de proteínas.

Umbral	Small	Human	Thermotoga	Thermus	Archaea	GFP	AFP	PLA2	Prion	BMP
IPE<0,01	0	0	0	0	0	0	89	0	319	0
	0	0	0	0	0	0	7	0	25	0
IPE<0,1	0	0	0	0	0	502	899	801	973	58
	0	0	0	0	0	39	70	62	76	5

En la tabla anterior, puede verse que el número de tripletas con IPE menor que 0.01, que es un valor bastante bajo de IPE. Sin embargo, encontramos que las proteínas humanas, las proteínas clasificadas por SCOP como “pequeñas”, las de Thermotoga, las de Thermus, las de Archaea, las GFP y las BMPs, no presentan en ninguna de sus tripletas, valores tan bajos. Sin embargo, las proteínas menos estables en su estructura secundaria, sí presentan un pequeño porcentaje de estas tripletas.

Pero cuando elevamos el umbral, para considerar las tripletas con IPE menor que 0.01, es decir, un IPE que habla de poca estabilidad estructural de la estructura secundaria (entendida como las tripletas que pueden aparecer en varias estructuras secundarias, con valores de frecuencia aproximadamente iguales. En este caso, las proteínas GFP, AFP, PLA2, Prión y BMP, presentan contenidos de estas tripletas más altos. Es llamativo el caso del prión, que está constituido en más del 75% por tripletas de baja preferencia estructural. Esto puede estar explicando la tendencia de esas proteínas, a cambiar el plegamiento de su estructura y alcanzar otras estructuras secundarias.

Lo presentado anteriormente, puede verse en la figura 4.3.2.3., que muestra a las proteínas del Prión, constituidas por las tripletas con IPEs más bajos, mientras que las de Humanos, las pequeñas y las de Thermotoga y Thermus, constituidas por tripletas con IPE más altos.

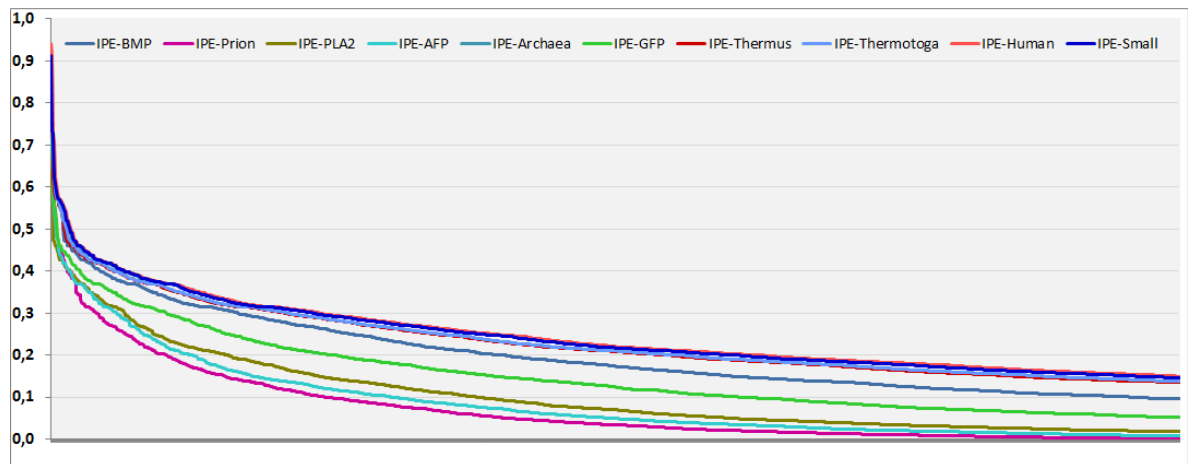


Figura Comparación entre los valores de IPE de las tripletas que componen a diferentes grupos de proteínas (TGFP-BMP, Prión, PLA2, AFP, Archaea, GFP, Thermus, Thermotoga, Human, Small). En el eje X, están las tripletas que componen a cada una de las proteínas, y no son las mismas entre unas y otras proteínas. Se representan los valores de IPE de las tripletas que componen a varios tipos de proteínas.

Construimos un archivo Excel que recibe la secuencia de aminoácidos de una proteína y puede calcular y contar el número de tripletas con alto ó bajo IPE, de manera que se puedan identificar en la secuencia proteica, zonas con valores altos ó bajos de IPE, identificando de esta manera, zonas más ó menos estables en su estructura secundaria.

Consideramos que este enfoque puede ser usado para analizar las proteínas de un proteoma, para identificar aquellas con menor tendencia a la estabilidad estructural y por tanto, más propensas a cambiar su plegamiento.

CONCLUSIONES CAPÍTULO 5 - PROPENSIONES, ENTROPÍA E IPE PARA AMINOÁCIDOS Y TRIPLETS

En este capítulo analizamos tres diferentes indicadores: Propensión, Entropía conformacional, e Indicador de Preferencia estructural desarrollado por nosotros, y los comparamos en términos del aporte al análisis de estructura secundaria, y la inferencia que se puede derivar de la misma. Estos tres indicadores se analizaron en aminoácidos individuales y también en triplets de aminoácidos.

Los valores de Chou y Fasman calculados tanto en aminoácidos individuales como en triplets varían significativamente entre diferentes grupos de proteínas, lo que significa que su capacidad de predicción es limitada al no considerar las particularidades de los ambientes de plegamiento y función de proteínas particulares. Por eso, su valor predictivo disminuye si se usan propensiones globales. Las propensiones deberían calcularse para grupos de proteínas con características similares, en condiciones similares, de manera que se maximice su valor predictivo. De aquí que resulta importante extender el concepto de propensión a triplets de aminoácidos.

La entropía conformacional resulta muy similar al inverso del Indicador de preferencia estructural (IPE) que proponemos, a nivel de aminoácidos individuales, pero no ofrece información sobre la estructura secundaria que un residuo tiende más a formar, como sí lo hace el IPE.

Cuando se evalúa en triplets de residuos de aminoácidos, la Entropía conformacional presenta un comportamiento diferente al del IPE, lo que muestra que el Indicador de Preferencia Estructural no es reemplazable por la entropía. La entropía mide la diversidad con que una tripleta aparece en diferentes estructuras secundarias; mientras que el IPE se enfoca más en determinar la dominancia con que una tripleta de aminoácidos, forma una estructura secundaria particular.

El Indicador de Preferencia Estructural (IPE) que desarrollamos en el marco de este proyecto, permite identificar secuencias cortas que se conservan en una estructura secundaria a lo largo de diferentes proteínas, y también identifica secuencias cortas que pueden cambiar su estructura secundaria de una proteína a otra, y que dependen del contexto de aminoácidos que le flanquean en la secuencia. Extendiendo este concepto de análisis de secuencias cortas, a la totalidad de la secuencia proteica, el IPE permite identificar las proteínas cuya estructura secundaria es estable, e indica cuándo una proteína está compuesta por triplets de aminoácidos que le confieren inestabilidad a la estructura secundaria. Por tanto, el IPE se puede usar en Ingeniería de proteínas, en que se requiera construir una estructura secundaria estable ó tendiente a cambiar.

El IPE permite también identificar triplets que no presentan tendencias a formar estructuras secundarias específicas, y que dependen del contexto de los residuos de aminoácidos que le rodeen. A estas triplets las hemos llamado “Triplets camaleón”, pues alcanzarían su estructura secundaria, dependiendo del entorno de secuencia y el ambiente que en se encuentren.

Se desarrolló una herramienta de Excel para el análisis de una secuencia de aminoácidos, en términos de la preferencia estructural, estimada con el IPE, que puede ser útil en el rediseño de proteínas.

Las tripleta WMC (Tryptófano-Metionina-Cisteína), que aparece predominantemente en hélices (HHH), y la tripleta HHH (Histidina-Histidina-Histidina), que aparece predominantemente en Giro al azar (UUU), son las que presentan los mayores valores de IPE, por encima de 0.94, lo que significa que cada una de estas tripletas, es claramente formadora de una estructura secundaria en particular. Consideramos que estas combinaciones de aminoácidos, son las más determinantes para la formación de una estructura secundaria. Estas tripletas se vieron en el capítulo anterior, como tripletas que aparecen con frecuencias muy altas en las estructuras secundarias que se mencionan aquí.

Algunas tripletas, principalmente constituidas por repeticiones de los aminoácidos de Cisteínas, Metioninas, Tryptófano, y Aspártico, resultaron ser tripletas con valor de IPE, cero, con lo que las consideramos tripletas camaleón, que se acomodan a la estructura secundaria de los aminoácidos del contexto que rodea la tripleta. Los residuos de Metionina y Tryptófano se presentaron entre los aminoácidos con bajas frecuencias de aparición bajas y aproximadamente iguales en todas las estructuras secundarias, según el capítulo 3; lo que valida la capacidad del IPE de categorizar secuencias cortas de acuerdo con su preferencia estructural.

BIBLIOGRAFÍA CAPÍTULO 5 – PROPENSIONES C&F, ENTROPIA CONFORMACIONAL E INDICADOR DE PREFERENCIA ESTRUCTURAL

Altschul, S.F. (1991). **Amino acid substitutions matrices from an information theoretic perspective.** J. Mol. Biol. 219, 555-665.

Anfinsen, C.B., Redfield, R.R., Choate, W.I., Page, J., y Carroll, W.R. (1954). **Studies on the gross structure, cross-linkages, and terminal sequences in Ribonuclease.** Journal of Biological Chemistry 207, 1: 201-210.

Anishetty, S., Pennathur, G. y Annishetty, R. (2002). **Tripeptide analysis of protein structures.** BMC Structural Biology. 2:9-16.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L., Studholme, D.J., Yeats C. y Eddy S.R. (2004). **The Pfam Protein Families Database.** Nucleic Acids Research, 32:D138-D141.

Bateman, A. Birney, E., Cerruti, L. Durbin, R. Etwiller, L. Eddy, S.R. Griffiths-Jones, S. Howe, K.L. Marshall, M. y Sonnhammer, E.L. (2002). **The Pfam Protein Families Database.** Nucleic Acids Research 30(1):276-280.

Bateman, A. Birney, E., Durbin, R., Eddy, S.R., Howe K.L. y Sonnhammer E.L. (2000). **The Pfam Protein Families Database:** Nucleic Acids Research 28:263-266.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R. Finn R.D. y Sonnhammer, E.L.L. (1999). **Pfam 3.1: 1313 multiple alignments match the majority of proteins.** Nucleic Acids Research 27:260-262.

Berman, H., Westbrook, J., Feng, Z., Guilliand, G., Bhat, T.N., Weissing, H., Shindyalov, I.N., y Bourne, P.E. (2000). **The Protein Data Bank.** Nucleic Acids Research. 28 (1) : 235-242.

Berman, H.M., Henrick, K. y Nakamura, H. (2003). **Announcing the worldwide Protein Data Bank.** Nat Struct Biol, 10: 98.

Carpino, L. A. y Han, G. Y. (1972). **The 9-fluorenylmethoxycarbonyl aminoprotecting group.** J. Org. Chem. 37, 3404–3409.

Chou K.C. (2000) **Prediction of tight turns and their types in proteins.** Anal Biochem 286: 1–16.

Chou, P.Y., y Fasman, G.D. (1974). **Prediction of protein conformation.** Biochemistry—USA 13:222-45.

Chou, P.Y., y Fasman, G.D.(1974). **Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins.** Biochemistry—USA 13:22-22. 1974.

Chou, PY y Fasman, GD. (1974). **Prediction of protein conformation**. Biochemistry, 13, 222-245. El artículo clásico de Chou y Fasman. El que armó el alboroto. Buscar este artículo, pues allí se explica cómo se obtuvieron los parámetros; a partir de qué fórmulas.

Chou, PY. (1989) En Fasman, GD. (ed.), **Prediction of protein structure and the principles of protein conformation : Prediction of protein structural classes from amino acid compositions**. New York, pp. 549-586. Artículo en el que recalculan los parámetros de predicción de cada aminoácido, usando un número mayor de proteínas (66).

Chou PY, Fasman GD. (1977). **Prediction of protein conformation: secondary structure**. Journal of Molecular Biology., 115, 135-175.

Coin, L., Bateman A. and R. Durbin. (2003). **Enhanced protein domain discovery by using language modeling techniques from speech recognition**. Proc. Natl. Acad. Sci. USA. 100(8):4516-4520.

Costantini, S., Colonna, G., Facchiano, A.M. (2006). **Amino acid propensities for secondary structures are influenced by the protein structural class**. Biochemical and Biophysical Research Communications 342, 441-451.

Chuan C.H. (2012). **Proteomics in Clinical Studies and its Applications**. Single Cell Biology: 1-2.

Djikaev, Y. S. y Ruckenstein, E. **Model for the Nucleation Mechanism of Protein Holding**. Journal of Physical Chemistry B, 111 (4), 886 -897, 2007.

Dill, K. A., Bromberg, S., Yue, K. Z., Fiebig, K. M. Yee, D. P., Thomas, P. D. y Chan, H. S. (1995). **Principles of protein folding - A perspective from simple exact models**. Protein Science., 4, 561-602.

Doig, A.J. y Baldwin, R.L. 1995. **N- and C-capping preferences for all 20 amino acids in alpha-helical peptides**. Protein Sci. 4(7): 1325–1336.

Epstein, H.F., Schechter, A.N., Chen, R.F., y Anfinsen, C.B. (1971). **Folding of staphylococcal nuclease: Kinetic studies of two processes in acid renaturation**. Journal of Molecular Biology, Volumen 60 (3): 499-508.

Fasman, G.D., 1989. Libro: **Prediction of Protein Structure and the Principles of Protein Conformation**. Pag 221. (ISBN 0306431319) New York (Plenum).

Fasman, G.D., (1987). Biochemistry, sección : **“This week’s citation classic”**. 30 : 32 : 21.

Kabsch W, Sander C. (1983). **Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features**. Biopolymers, 22(12):2577-637.

Kruus, E., Thumfort, P., Tang, C., Wingreen, N.S. 2005. **Gibbs sampling and helix-cap motifs**. Nucl. Ac. Res. 33(16):5343-5353.

Mejía, D.F, Tischer, I. (2007). **Design of a new indicator of structural preference for residue triplets into proteins**. Capítulo del libro: “Proceedings, International Seminar on Genomics,

Proteomics, Bioinformatics and Systems Biology” (ISBN 978-958-9451-21-2), pags 123-130. Popayán, Colombia.

Paskowskia, D.J., Stevensb, E.S., Bonorac, G.M., Tonioloc, C. (1978) **Beta sheet formation by L-methionine oligopeptides**. Biochimica et Biophysica Acta (BBA) - Protein Structure 535 (2):188–192

Smith, L.J., Fiebig, K.M., Schwalbe, H., Dobson, C.M. (1996). **The concept of a random coil: Residual structure in peptides and denatured proteins**. Folding & Design. 1:R95–R106.

Strait, B.J. y Dewey, T.G. (1996). **The Shannon information entropy of protein sequences**. Biophysical Journal. 71(1): 148–155.

CAPÍTULO 6. REDISEÑO DE BMPs Y B-LACTAMASAS MEDIANTE RECONSTRUCCIÓN ANCESTRAL

Entre 2010 y 2012, hice mi pasantía doctoral, entre los laboratorios de Química Física e Ingeniería de Proteínas de la Universidad de Granada, y el laboratorio de Biología Sintética del Instituto Tecnológico de Georgia, GeorgiaTech. Durante ese tiempo, reconstruimos computacionalmente, Proteínas Morfogenéticas de Hueso, para su uso en cirugías ortopédicas y maxilofaciales, y enzimas Beta-Lactamasas para su uso en la evaluación de nuevas moléculas, candidatas a ser antibióticos de nuevas generaciones. Debido a lo anterior, este capítulo se presenta como un capítulo independiente, con su propia tabla de contenido, introducción, etc.

En el presente capítulo, se presentan los resultados de este trabajo. Se ha incluido una tabla de contenido a este capítulo, en forma particular, dado el carácter independiente de este trabajo, con respecto al trabajo de los capítulos anteriores; por ello, este trabajo presenta sus propios objetivos, introducción, resultados y conclusiones.

TABLA DE CONTENIDO

TABLA DE CONTENIDO	172
INTRODUCCIÓN	173
LA INGENIERÍA DE PROTEÍNAS	173
Biología Sintética Evolutiva y Reconstrucción de las secuencias ancestrales	174
LAS PROTEÍNAS MORFOGENÉTICAS DE HUESO	175
Las BMP, regeneración ósea, efectos de la des-regulación, y proteínas recombinantes	175
Los problemas óseos y el impacto clínico de las BMP	177
Importancia y un poco de historia	178
Sobre la estructura de las BMPs	180
Mecanismo de acción de las BMP	184
OBJETIVOS	187
OBJETIVO GENERAL	187
OBJETIVOS ESPECÍFICOS.....	187
MATERIALES Y MÉTODOS.....	188
SELECCIÓN DE LA SECUENCIA DE INICIO (SECUENCIA “QUERY”)	188
BÚSQUEDA DE SECUENCIAS EN LA BASE DE DATOS DEL NCBI	188
ALINEAMIENTO DE LAS SECUENCIAS SELECCIONADAS	188
ARBOLES FILOGENÉTICOS DE LAS SECUENCIAS SELECCIONADAS	188
RECONSTRUCCIÓN DE LAS SECUENCIAS ANCESTRALES DE LOS TGF-B	189
RESULTADOS Y DISCUSIÓN.....	190
SIGUIENTES PASOS Y PERSPECTIVAS FUTURAS	204
BIBLIOGRAFÍA	205
RESURRECCIÓN DE LACTAMASAS ANCESTRALES.....	189
ARTÍCULO DE RESURRECCIÓN DE LACTMASAS ANCESTRALES, PUBLICADO EN JACS.....	190

6.1. INTRODUCCIÓN

6.1.1. LA INGENIERÍA DE PROTEÍNAS

Podría decirse que las proteínas son las verdaderas protagonistas de la vida, desde el punto de vista molecular, dada la enorme gama de actividades y funciones que realizan en los seres vivos. Esto ha conducido a que el estudio de las proteínas sea un campo “caliente”, que genera muchas publicaciones cada año. Pero un campo de la ciencia alcanza su madurez, cuando el conocimiento acumulado adquirido permite hacer modificaciones y nuevos diseños para su uso y aplicación en otras áreas. En cuanto al campo específico de las proteínas, esto se conoce como “Ingeniería de proteínas”.

El término “Ingeniería de proteínas”, hace referencia a la modificación, diseño ó rediseño de proteínas con propiedades alteradas, ó incluso nuevas. La ingeniería de proteínas tiene un gran interés biotecnológico dada la amplia gama de posibles aplicaciones prácticas de las funciones proteicas. Por el momento, estas aplicaciones son fundamentalmente de tres tipos: aplicaciones biomédicas/terapéuticas (Lazar et al. 2003), aplicaciones tecnológicas/industriales (Kirk et al., 2002), y aplicaciones en nanotecnología (van den Heuvel y Dekker. 2007).

Las aplicaciones terapéuticas de proteínas se basan generalmente en anticuerpos monoclonales y en proteínas involucradas en comunicación celular (hormonas, factores de crecimiento, citoquinas, etc.) (Lazar et al. 2003). Las aplicaciones tecnológicas se basan esencialmente en proteínas con actividad enzimática e incluyen su uso en detergentes, industria alimentaria, alimentación animal, industria textil, síntesis orgánica, etc. (Kirk et al., 2002). Las aplicaciones de las proteínas en nanotecnología, hacen uso esencialmente de proteínas que actúan como motores moleculares (la bomba de protones, polimerasas, proteínas de la contracción muscular, etc.) (van den Heuvel y Dekker. 2007).

Usualmente, la ingeniería de proteínas hace uso de la introducción de mutaciones en la secuencia de aminoácidos. El punto crítico es introducir mutaciones que causen los cambios que se desean introducir en la estructura para modificar la estabilidad, la función ó alguna otra propiedad de la proteína. Para ello, hay dos aproximaciones que se podrían considerar clásicas:

1. El diseño racional (computacional), en que, conociendo la estructura de la proteína, y mediante un modelo teórico de cómo los cambios afectarían a las propiedades de la proteína, se predicen los efectos de las mutaciones.
2. Evolución dirigida *in vitro* (Tobin et al, 2000), en que se cuenta con una amplia biblioteca de variantes de proteínas, en el que se obtienen variantes con las características deseadas, mediante tamizaje y selección.

Estas dos aproximaciones tradicionales, aunque presentan algunas limitaciones, condujeron a muchas “historias exitosas”. El diseño racional está frecuentemente dificultado por el limitado

conocimiento de la relación estructura-energética de las proteínas mientras que la exitosa evolución dirigida *in vitro*, puede requerir del tamizado de bibliotecas muy grandes (lo cual puede no ser práctico, ni conveniente desde el punto de vista de la relación costo-beneficio).

Hay una “tercera aproximación”, que en los últimos años ha mostrado un incremento en el número de trabajos publicados que intentan usar información evolutiva en ingeniería de proteínas. La razón detrás del desarrollo de este enfoque es doble. En primer lugar, como las proteínas existentes hoy en los seres vivos, son el resultado de la selección natural, la información acerca su evolución puede usarse para guiar la ingeniería de proteínas. Como segundo punto, las bases de datos de secuencias están creciendo exponencialmente y contienen jugosa información evolutiva que puede ser aplicada de forma inmediata. De este enfoque evolutivo, tal vez el más conspicuo y con resultados exitosos, sea la Reconstrucción de secuencias ancestrales, que algunos autores consideran que pertenece al campo conocido como Biología sintética evolutiva.

Biología Sintética Evolutiva y Reconstrucción de las secuencias ancestrales

Es una metodología relativamente reciente, usada para inferir información evolutiva, ecológica, bioquímica, fenotípica, ambiental ó biogeográfica asociada con nodos ancestrales extintos en un árbol filogenético. La Reconstrucción de Secuencias Ancestrales (ASR, por su sigla en inglés), es la generación de secuencias de proteínas, genes, ó incluso genomas completos extintos, hipotéticamente correspondientes a tiempos antiguos, a partir de las secuencias de proteínas, genes ó genomas existentes hoy.

Originalmente, el famoso químico ganador de dos premios Nóbel y padre de la química estructural, Linus Pauling y uno de los padres de la Evolución molecular Emile Zuckerklandl, quienes propusieron el modelo del reloj molecular, propusieron la reconstrucción de proteínas antiguas extintas (Pauling y Zuckerklandl, 1963). Pocos años después, Margaret Dayhoff, posiblemente una de las madres de la Bioinformática, propuso que las secuencias actuales podrían dar información sobre secuencias del pasado (Eck y Dayhoff, 1966). A partir de esta propuesta, Walter Fitch, de la Universidad de California en Irvine, propuso el enfoque de mínimo cambio (Máxima Parsimonia) para definir el proceso de cambio evolutivo en un árbol (Fitch, 1971).

Esto llenó un vacío en la posibilidad de ir hacia atrás en el pasado, dentro de una filogenia y más de 45 años después de la propuesta de Pauling y Zuckerklandl, hasta que dos investigadores estadounidenses retomaron el concepto y reconstruyeron una enzima Ribonucleasa (Benner y Allemann, 1989). Hasta aquí, el trabajo había sido básicamente matemático y computacional; sería Steven Benner, y su grupo de investigación, quienes llevarían al laboratorio las secuencias reconstruidas, en un proceso que se conoce como resurrección. A través de este estudio, se demostró que la reconstrucción ancestral podría dar como resultado proteínas funcionales con características particulares (Stackhouse, et al., 1990)

Entrando los 90, mediante un análisis de Máxima Parsimonia, se infirieron las primeras secuencias proteicas de especies extintas, para luego sintetizar los genes y expresar las moléculas (Malcolm et al., 1990; Adey et al., 1994; Jermann et al., 1995). Luego, los evolucionistas moleculares depuraron los modelos matemáticos y evolutivos que usan Máxima Parsimonia (Yang et al., 1994; Yang et al., 1995; Swofford y Maddison, 1987), de Máxima Verosimilitud (Koshi y Goldstein, 1996), y de Distancia (Pupko et al., 2000), para inferir secuencias ancestrales a partir de secuencias actuales.

Posteriormente se reportaron reconstrucciones comparativas entre diferentes modelos matemáticos (Zhang y Nei, 1997). Posteriormente se estudiaron las propiedades fisicoquímicas de proteínas resucitadas, encontrando propiedades tan interesantes como termoestabilidad aumentada en algunas de ellas (Malcolm et al., 1990; Stewart, 1995). Con esto, se ha inferido información acerca de los ambientes antiguos del planeta (Gaucher et al., 2003), estabilidad y actividad de proteínas extintas (Thornton, J. 2004), y hasta aspectos fisiológicos y etológicos como si los dinosaurios cazaban de noche (Chang et al., 2002; Chang, 2003).

La reconstrucción de secuencias ancestrales también ha permitido deducir el camino evolutivo de algunos genes y proteínas, con lo que se pueden identificar los cambios específicos en un aminoácido ó nucleótido, que debieron ocasionar cambios funcionales en la proteína (Stewart et al. 1987; Swanson et al. 1991), También se han podido hacer inferencias evolutivas sobre diferentes organismos (Ugalde et al., 2004).

El subcampo de las reconstrucciones de secuencias ancestrales reúne un pequeño grupo de expertos en el mundo, que han ofrecido depuraciones a las metodologías y ofrecido múltiples aplicaciones de las mismas (Liberles, 2007). Hoy en día, se ha estado aplicando a la construcción de proteínas con propiedades particulares, para Ingeniería de Proteínas, con resultados (Libertini y di Donato, 1994; Purdy, 2008; Dean y Golding, 1997). Nuestro grupo de investigación, en colaboración con un grupo de investigación en GeorgiaTech, Atlanta, y la Universidad de Columbia en los Estados Unidos, ha participado en la reconstrucción, resurrección y evaluación experimental de enzimas Tiorredoxinas (Perez-Jimenez, et al, 2011).

El presente trabajo forma parte de las primeras reconstrucciones de proteínas ancestrales para su evaluación en el laboratorio, realizada en nuestro grupo de investigación. Hemos reconstruido proteínas ancestrales de BMP2, BMP4, y otras proteínas de la superfamilia de los TGF- β , que participan en procesos de desarrollo en el organismo de vertebrados. Dentro de nuestra reconstrucción, hemos obtenido secuencias proteicas ancestrales de nodos de más de 750 millones de años en el pasado.

Para el presente trabajo, hemos trabajado en colaboración con el grupo de investigación Biología Sintética Evolutiva liderado por Eric Gaucher, en GeorgiaTech, conocido experto en la reconstrucción de secuencias ancestrales. Esta investigación se enmarca dentro de las prioridades del Consorcio científico para la Ingeniería de Proteínas (Prodestech por “Protein Design and Technologies”), que abarca varios de los grupos que trabajan en punta en este campo en España, y forma parte de un proyecto colaborativo con el grupo de investigación en Ingeniería de tejidos del Instituto de Estudios Biológicos (IEB), de la Universidad Complutense de Madrid, liderado por el Dr. José Luis López Lacomba (<http://prodestech.cib.csic.es/members/group-5>, <http://portal.ucm.es/web/ingenieria-de-tejidos/inicio>), que trabaja en el desarrollo de aplicaciones terapéuticas de las Proteínas Morfogenéticas de Hueso, y las estrategias para transferir los desarrollos a la clínica y la industria.

LAS PROTEÍNAS MORFOGENÉTICAS DE HUESO

Las BMP, regeneración ósea, efectos de la des-regulación, y proteínas recombinantes

Las BMPs se consideran un grupo primordial de señales morfogénicas, inductoras de la formación de hueso y cartílago. Su función como señales celulares se ha hecho evidente cuando

ocurre una regulación inadecuada de las señales, causando procesos patogénicos (Bleuming et al., 2007). En especial, el cáncer implica una regulación incorrecta del sistema de señales de BMPs. A manera de ejemplo, la ausencia de señalización, es un factor importante en la progresión del cáncer de colon (Kodach et al., 2008), y la sobre-activación de la señalización debida a una esofagitis inducida por reflujo, provoca el síndrome conocido como Esófago de Barret, que es crítico en el desarrollo del adenocarcinoma en la porción terminal del tracto gastrointestinal (Milano et al., 2007).

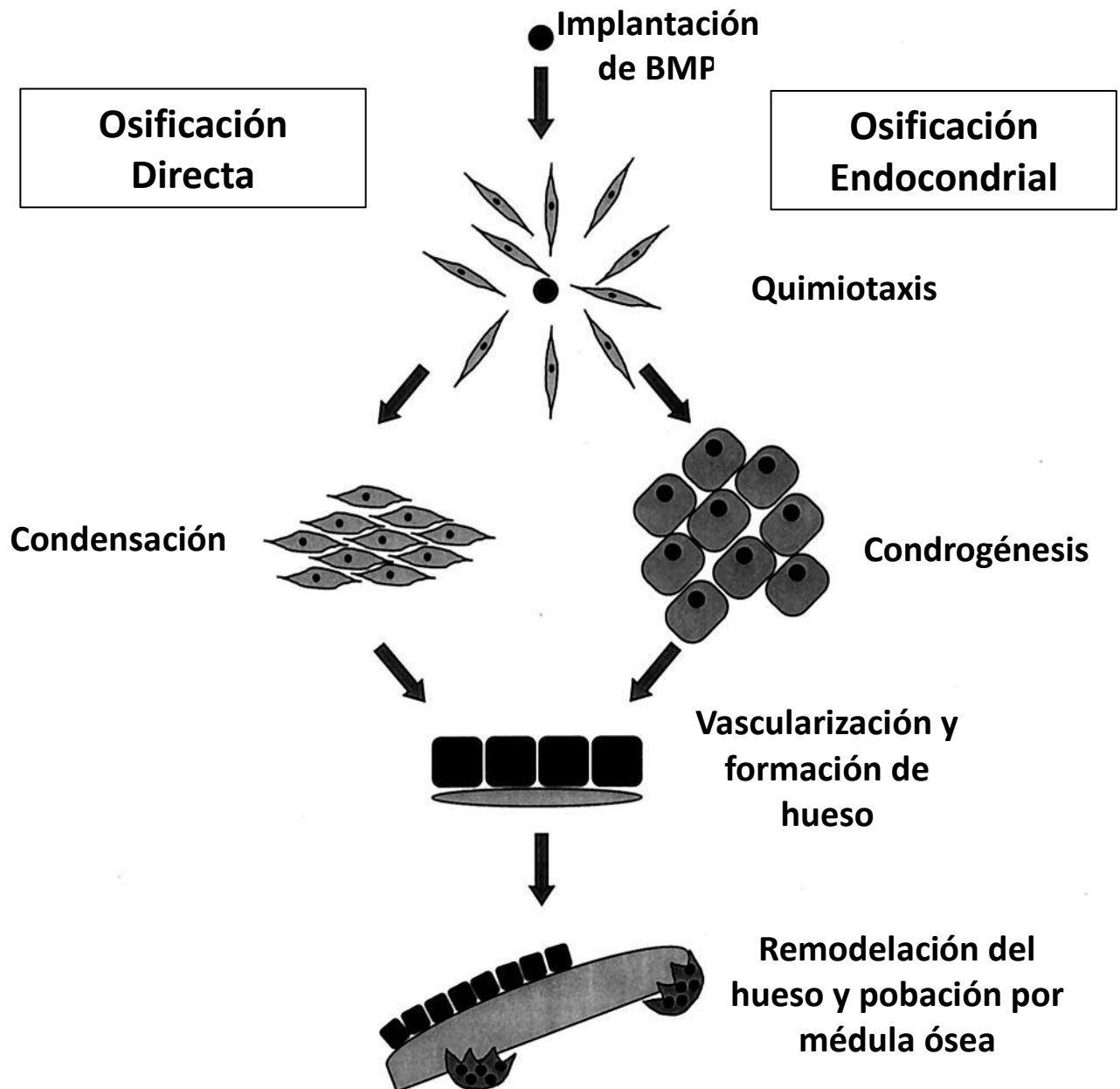


Figura 5.1. Eventos celulares que ocurren después de la implantación de las BMP. Estas inducen la formación de hueso endocondrial (mediante un intermediario de cartílago), y directo (intramembranoso). El resultado final en cada caso, es tejido óseo que se remodela y se llena con médula ósea hematopoiética.

Dado su papel en la regeneración de tejido óseo, las BMPs se han empezado a usar terapéuticamente en cirugías ortopédicas y maxilofaciales (Wikesjö et al., 2009, Moghadam et al., 2001). En medicina regenerativa, las BMPs se dirigen al sitio de la fractura para su incorporación en un implante óseo, que las libera gradualmente a lo largo de varias semanas, para permitir la formación de hueso a medida que ocurre la estimulación del crecimiento por las BMPs. Hace pocos años, la FDA aprobó dos productos de BMP para aplicaciones clínicas (principalmente para tratar las fracturas de huesos largos y regeneración de discos intervertebrales), mediante su liberación en una matriz de colágeno que se implanta en el sitio de la fractura. Uno de los productos aprobados es conocido como “Infuse bone graft” de la compañía Medtronic, aprobado en 2002 para su uso en la fusión espinal y hoy también en aplicaciones dentales (información sobre la patente, en http://www.accessdata.fda.gov/cdrh_docs/pdf/P000058a.pdf). El otro producto aprobado es conocido como “OP-1 BMP-7” de la compañía Stryker Biotech, y se usa como ayuda terapéutica en la fusión de vértebras para prevenir el trauma neurológico, y para el tratamiento para curar una tibia fracturada donde haya fallado un injerto óseo como tratamiento previo (Zimmermann et al., 2007; http://www.healthpointcapital.com/research/2004/04/12/fda_approves_second_growth_factor_for_the_spine/).

Aparte de estos dispositivos médicos aprobados por la FDA, hay varias patentes relacionadas con las proteínas morfogenéticas de hueso, de las que se han identificado, al día de hoy, 23 tipos de BMPs en el ser humano, de las cuales, la que al parecer juega el papel más crítico en el desarrollo óseo y que es posiblemente la que ha recibido mayor atención y sobre la que se realiza más investigación, es la BMP-2, de la que Wozney tiene una patente por una versión recombinante de la proteína (Oficina de patentes de los Estados Unidos: <http://patft.uspto.gov/netacgi/nph-Parser?Sect2=PTO1&Sect2=HITOFF&p=1&u=/netahtml/PTO/search-bool.html&r=1&f=G&l=50&d=PALL&RefSrch=yes&Query=PN/5618924>). Esta patente, con número de registro de patente en los Estados Unidos US5618924, otorgada el 8 de Abril de 1997, y aunque expiraba el 8 de Abril de 2014, obtuvo una extensión de la patente hasta el 2016 (Medtronic, Annual Report, 2009). Considerando el costo del tratamiento con las proteínas patentadas y aprobadas, y la actividad de la BMP-2 comercial, es de gran interés diseñar una BMP-2 con actividad y estabilidad magnificadas, y a un costo de tratamiento más razonable.

Los problemas óseos y el impacto clínico de las BMP

En un estudio presentado en el Congreso Europeo de Osteoporosis y Osteoartritis, realizado en Valencia, en Marzo de 2011, por el presidente de la International Osteoporosis Foundation, se mostró que los costos de las fracturas por fragilidad ósea en Francia, Alemania, Italia, España, Suecia y el Reino Unido, ascienden a 31 billones de Euros al año (Ström *et al.* 2011). También se han hecho estudios sobre el costo social y económico de las fracturas derivadas de osteoporosis en España, mostrando que sólo por fracturas de cadera, se presentan más de 40.000 casos al año, más otras fracturas relacionadas con la osteoporosis (Solán et al., 2010; Herrera et al., 2006; Álvarez-Nebreda, 2008)

Las fracturas traumáticas implican un factor importante de enfermedad. De acuerdo con la Organización Mundial de la Salud (OMS), los accidentes traumáticos implican unos 500 millones de

dólares al año en el mundo, de los cuales una parte importante implica fracturas y cirugías óseas (Hak, 2005, World Health Organization, 2003).

Los primeros estudios clínicos se iniciaron con 11 pacientes, que mostraron resultados prometedores en términos de la regeneración ósea (Howell et al. 1997). Hoy en día, se han hecho estudios clínicos con BMP-2 para el tratamiento de varias patologías óseas. El éxito comercial de la BMP-2 recombinante humana ha estimulado la evaluación clínica en otros subconjuntos de defectos óseos y la investigación en nuevos factores de crecimiento. Los productos se han aprobado también para su uso en procedimientos de implantación dental (Cochran et al., 2000). El uso de los factores de crecimiento para reparación de cartílago, tendón ó tejidos blandos, aún están en estados preclínicos, constituyéndose en un nicho con ventajas competitivas para quienes investigan en ello.

La BMP-2 se aprobó para la fusión espinal y para fracturas de tibia (Govender et al., 2002) y se ha evaluado su efectividad en la reparación de fracturas (Eihorn et al., 2003); la BMP-7 (también conocida como Proteína Osteogénica-1 (OP-1), recibió una exención humanitaria como dispositivo para uso en fracturas no articulares que han fallado en curarse por otros tratamientos (Friedlaender et al., 2001). Las BMP-2 y BMP-7, aunque comerciales y aprobadas por la FDA para su uso clínico, tienen un inconveniente importante: su alto costo, que puede oscilar entre 5000 y 8000 euros en un tratamiento convencional (Hak, 2005).

Producir proteínas recombinantes de BMPs, ofrece perspectivas interesantes en términos de su utilidad terapéutica en tratamientos para la regeneración ósea, pero también en otros aspectos en que esté implicada la regeneración de tejidos y el control de la misma mediante señalización celular.

Importancia y un poco de historia

Las enfermedades óseas ó relacionadas con articulaciones y huesos, los eventos traumáticos, las deformidades del desarrollo y la resección de tumores, son causa frecuente de defectos óseos que pueden afectar seriamente la calidad de vida y comprometer las habilidades quirúrgicas de cirujanos ortopédicos y maxilofaciales, y tienen impacto en los costos asociados con el sistema de salud (Kenley et al., 1993).

Es común que la reparación de deficiencias óseas implique técnicas quirúrgicas como injertos autólogos, aloinjertos, dispositivos de fijación interna ó externa, estimulación eléctrica, e implantes aloplásticos (Kalfas, 2001). Se estima que sólo en los Estados Unidos, se realizan más de 500.000 procedimientos de injertos óseos, la mitad aproximadamente, dedicados a la fusión espinal (Greenwald et al., 2010). La tecnología existente, aunque efectiva en muchos casos, aún presenta dificultades y desventajas. Desde hace algunos años, la Biotecnología ofrece alternativas a la reparación ósea, mediante la administración de factores de crecimiento, con la consecuente mejora en la calidad de vida de los pacientes, disminución de costos de tratamientos y cuidados, y reparaciones permanentes sin efectos colaterales importantes (Kenley et al., 1993; Murray-Rust, 1993; Nandi, 2010).

Los factores de crecimiento implicados en el desarrollo y reparación del hueso, son las Proteínas Morfogenéticas de Hueso (BMPs por su sigla en inglés). Las BMPs promueven la formación de hueso y cartílago, y participan en la regeneración de los mismos en el caso de heridas y fracturas. Estas proteínas son parte de una superfamilia de proteínas conocida como Factor de Crecimiento Transformante tipo β (TGF- β por su sigla en inglés), que son citoquinas reguladoras de procesos celulares, que incluyen proteínas participantes en control del crecimiento celular, la proliferación celular y la apoptosis, entre otras (Burt, 1992; Herpin, et al., 2004).

Los estudios sobre regeneración de hueso empiezan en 1889, cuando Senn observó que el hueso descalcificado podía inducir la curación de defectos óseos (Senn, 1889). La primera evidencia de la formación de hueso ectópico, la dio Lavender, cuando inyectó extractos de hueso crudo en músculo (Lavender, 1934, 1938). Lacroix en 1945 sugirió que la regeneración de hueso debería estar dirigida por los mismos factores implicados en el crecimiento (Lacroix, 1945). Ya en 1965, el científico y ortopedista estadounidense, Marshall Urist, en un trabajo pionero reportó el descubrimiento de una familia de proteínas en la médula ósea, que tienen capacidad osteoinductiva, cuando usó un segmento de hueso desmineralizado y liofilizado, que indujo la formación de nuevo hueso cuando fue implantado en músculo de conejo (Science, 1965).

Con esto, Urist descubrió que el compuesto activo responsable de la regeneración ósea era una mezcla de proteínas, a las que bautizó con el nombre de “Proteínas morfogenéticas de hueso” (Urist y Strates, 1971); la osteoinducción es la habilidad de un material de promover la formación y el crecimiento de hueso, sobre la superficie de hueso ya existente, también se define como la formación de hueso *de novo* en tejidos duros ó blandos (Wozney, 2002), con claras ventajas para la reconstrucción biológica con fines terapéuticos, reemplazando los trasplantes óseos, con procedimientos donde se induce formación de hueso del mismo paciente.

La formación de hueso es un evento celular que implica una cascada de múltiples pasos, de los cuales las etapas clave son: quimiotaxis, mitosis y diferenciación (Reddi y Huggins, 1972). Estos estudios despertaron interés en la comunidad científica y se llegó a investigar más exhaustivamente en la purificación de estas proteínas, ya que la matriz de hueso desmineralizado es insoluble, con lo que un avance importante fue precisamente el uso de agentes caotrópicos como 4M Hidrocloruro de Guanidina, 8M Urea, ó 1% SDS para ello; con esto, fue posible crear un bioensayo reproducible para la formación de hueso ectópico (Sampath y Reddi, 1981). El ensayo se basaba en la actividad de la enzima Fosfatasa alcalina y en el contenido de calcio en el hueso recién formado. Reddi propuso que las BMPs eran responsables del inicio de la cascada de eventos en que las células progenitoras en la médula ósea, se ven inducidas por estas proteínas para producir osteocitos que conducen a la formación de hueso (Reddi y Huggins, 1972; Reddi, 1981).

Sin embargo, el extracto sólo ó el residuo insoluble por sí mismos no inducen la formación de hueso, lo que sugiere que en la célula se necesita un conjunto de eventos coordinados que implican tanto al extracto soluble como al sustrato de colágeno insoluble (Sampath et al, 1987; Luyten et al., 1989). Wozney logró clonar las BMPs que mostraron ser biológicamente muy activas, lo que significó un salto en el desarrollo del campo de investigación de estas proteínas (Wozney et al., 1988), y condujo a muchos avances basados en el uso de BMPs recombinantes para la aplicaciones clínicas como fusión espinal, curación de fracturas e ingeniería de tejidos dentales (Nakashima y Reddi, 2003; Seeherman y Wozney, 2005). El grupo de Reddi depuró la purificación y secuenció parcialmente la primera proteína osteogénica en 1989 (Luyten, 1989).

Las BMPs humanas se producen ahora en grandes cantidades mediante tecnologías recombinantes. En 2002, la Administración de Alimentos y Drogas de los Estados Unidos (FDA por "Food and Drug Administration") aprobó dos productos que contienen rhBMP-2 y rhBMP-7 en matrices de colágeno absorbible para la fusión espinal y la curación de fracturas de huesos largos (McKay et al., 2007; White et al., 2007). Reddi en 1997 y Wagner en 2010 propusieron cambiar el nombre de "bone morphogenetic proteins" a "body morphogenetic proteins" dada su participación en la señalización de otros procesos en el organismo, además del desarrollo de hueso (Reddi, 2005, Wagner, *et al*, 2010).

Sobre la estructura de las BMPs

Las BMPs se sintetizan como moléculas de preproteína. Antes de la dimerización, estas moléculas se cortan proteolíticamente en un sitio consenso de Arg-X-X-Arg, para generar las proteínas maduras que se dimerizan posteriormente. Se ha demostrado que la región N-terminal controla la estabilidad de la proteína madura procesada y que la secuencia corriente-abajo adyacente al corte, determina la eficiencia del mismo (Constam y Robertson, 1999).

La formación del dímero con un nudo de Cisteínas, es requisito para la función inductora de hueso. Las BMPs pueden ser activas como homodímeros y como heterodímeros. Su estructura primaria es similar en un 40 a 50% a la de los TGF- β , y como se mencionó antes, forman parte de esta superfamilia (Wozney et al., 1988). Aunque como las TGF- β , inducen la expresión de la Fosfatasa alcalina y la Osteocalcina (que actúan como marcadores de diferenciación), las TGF- β 1 inhiben dramáticamente la expresión de Osteocalcina y la actividad de la Fosfatasa alcalina (Spinella-Jaegle et al., 2001).

La resolución de la estructura tridimensional de la BMP-7 mediante cristalografía de rayos X, mostró que aunque hay una limitada identidad de secuencia, las BMP y las TGF- β 2 comparten un plegamiento común. Dentro de la subfamilia, hay mayor similaridad en secuencia: Mientras que la BMP-7 comparte menos del 38% de identidad en secuencia con la TGF- β 2, presenta una similaridad del 60 y del 85% con la BMP-2 y la BMP-5, respectivamente (Spinella-Jaegle et al., 2001; Griffith et al., 1996).

Los monómeros presentan 3 puentes de disulfuro intracatenarios que forman un nudo de Cisteína que se constituye en el núcleo del monómero, y un puente de disulfuro intercatenario con el otro monómero, que estabiliza el dímero activo (Figura 5.2). El monómero también presenta cuatro cadenas de hoja β antiparalela que forman dos proyecciones en forma de dedo hacia afuera del nudo (Figura 5.3), Tabas et al., 1991). El gen de la BMP-2 en humanos, se encuentra en el cromosoma 20, en la localización 20p12.3 (Tabas et al., 1991).

```
MVAGTRCLLALLLPQVLLGGAAGLVPELGRRKFAAASSGRPSSQPSDEVLSEFELRLLSMFGLKQRPTPSRDAV
VPPYMLDLYRRHSGQPGSPAPDHLRLERAASRANTVRSFHHEESLEELPETSGKTTRRRFFNLSSIPTTEEFITSAE
LQVFREQMQDALGNSSFHHRINIYEIIPATANSKFPVTRLRLVNQNASRWESFDVTPAVMRWTAQGHAN
HGFVVEVAHLEEKQGVSKRHVRISRLHQDEHSWSQIRPLLVTFGHDGKGHPLHKREKROAKHKQRKRLKSSC
KRHPLYVDFSDVGWNDWIVAPPGYHAFYCHGECPPFLADHLNSTNHAIVQTLVNSVNSKIPKACCVPTLSAIS
MLYLDENEKVVLKNYQDMVVEGCGCR
```

Figura 5.2. Secuencia de los 396 residuos de la preproteína de la BMP-2 humana. En color azul, la secuencia de los 114 residuos de la proteína madura activa.

La BMP-2 humana, que es la proteína blanco en el presente trabajo, se sintetiza como una larga proteína de 396 residuos, que luego se madura cortando la parte amino terminal, y dejando los últimos 114 residuos del extremo carboxilo terminal (Figura 5.2).

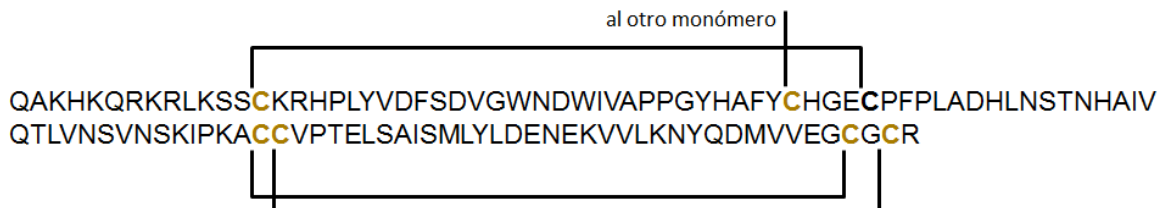


Figura 5.3. Secuencia de los 114 residuos de la BMP-2 humana madura, con las Cisteínas en color, y los enlaces de disulfuro intra e intercatenarios.

En la Figura 5.3, se pueden ver en la secuencia de la BMP-2 madura, las Cisteínas y los puentes de disulfuro que forman, tanto intracatenarios, como intercatenarios con el otro monómero para formar el dímero activo de la BMP-2. Los puentes de disulfuro forman un nudo que estabiliza al dímero activo (Figura 5.3).

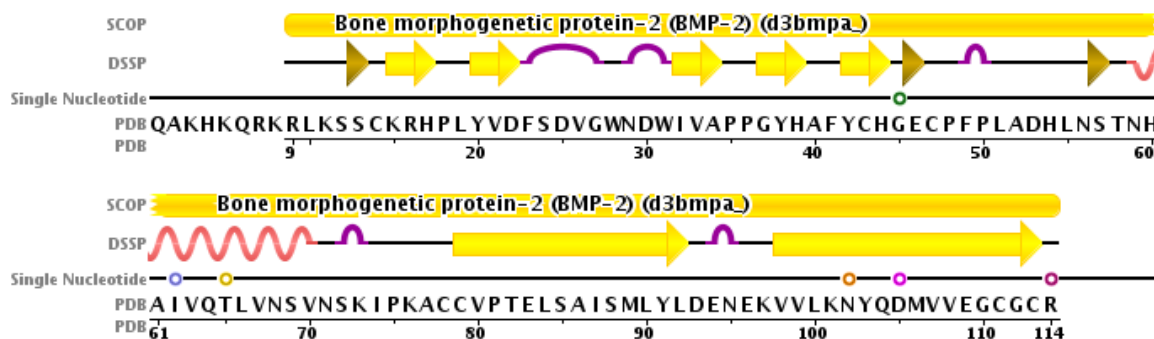


Figura 5.4. Topología de la BMP-2 humana (generada en www.pdb.org).

La BMP-2 presenta dos largas hélices hacia el extremo carboxilo, y cinco hélices cortas hacia el segundo tercio de la proteína, cerca del extremo amino. La molécula presenta un largo hilo Beta hacia la mitad de la proteína, entre los residuos 59 y 70 (Figura 5.4 y 5.5).

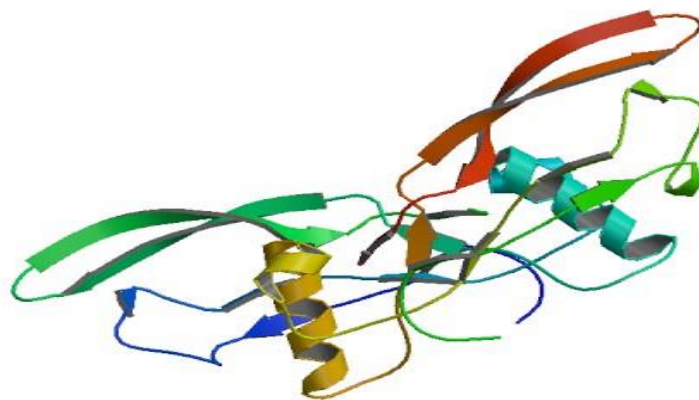


Figura 6.5. Estructura tridimensional del dímero de BMP-2 humana (3BPM).

La BMP-2 interactúa con su receptor, conocido como BRIA (por “BMP receptor IA”), mediante interacciones de van der Waals, y la formación transitoria de enlaces de hidrogeno. Los residuos críticos para la interacción, son Leu51 y Asp53 en la BMP-2. Se han construido variantes L51P que no se unen a los conocidos como receptores tipo I, como el BRIA, que confirman la importancia de este residuo para la interacción (Keller et al., 2004). Otros residuos importantes para la interacción de BMP-2 con su receptor, porque participan en la formación de enlaces de hidrógeno, son W28(NE1), L51(N y O), D53(OD2 y N), N59(OD1), S69(O y OG), Y103(OH) (lo que está entre paréntesis corresponde al átomo que participa en el enlace) (Keller et al., 2004). La estructura del complejo ligando-receptor se ve en las figuras 6.5 y 6.6.

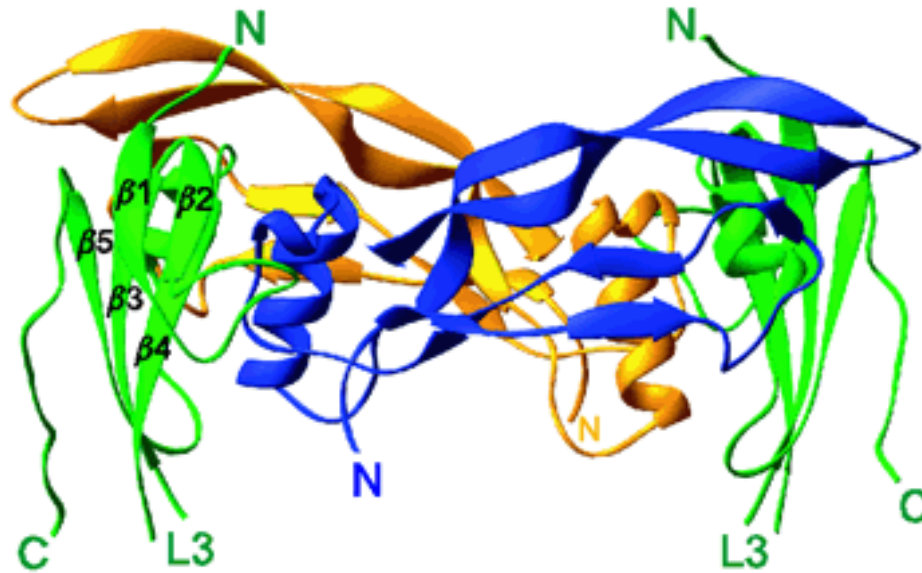


Figura 6.6. La estructura del complejo BMP-2-BRIAec. Vista lateral del lado próximo a la membrana. En naranja y azul, monómeros de BMP-2. En verde, el receptor BRIAec. Se etiquetaron los elementos de estructura secundaria extremos de la cadena y los giros 1 y 3 del receptor (tomada de Kirsch, T. Et al. 2000).

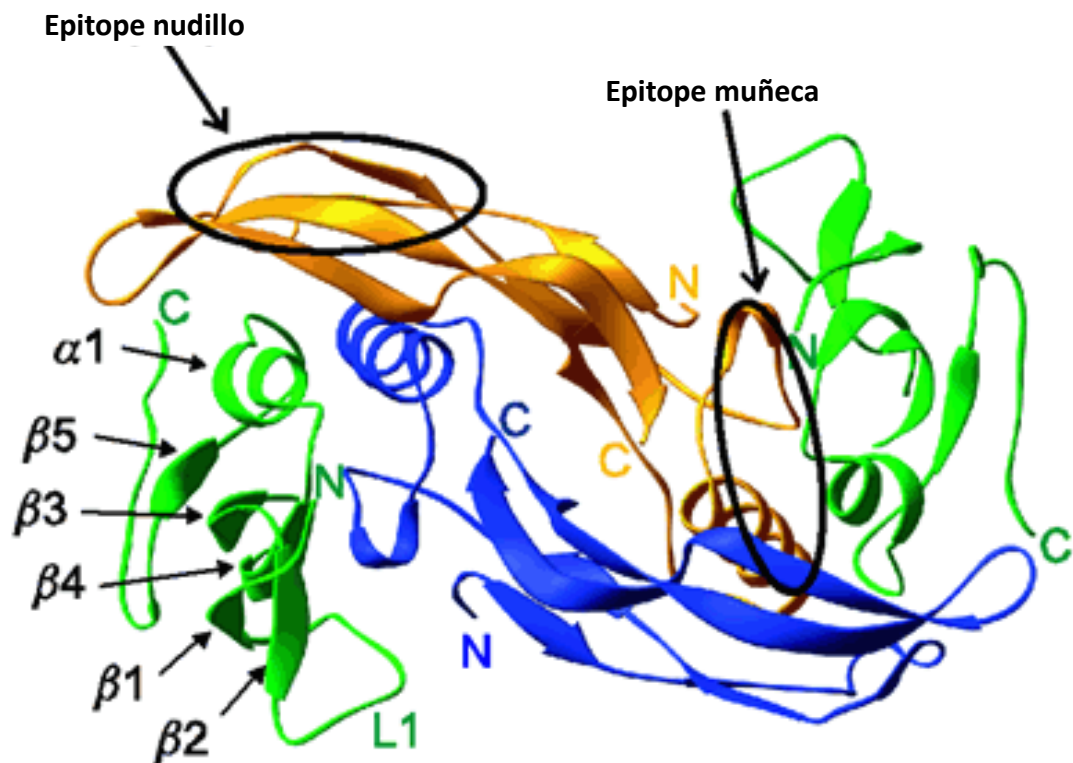


Figura 6.7. Estructura del complejo BMP-2-BRIAec. Vista desde arriba, aproximadamente a lo largo de eje bilateral del complejo. En dorado y azul, monómeros de BMP-2. En verde, el receptor BRIAec. Se etiquetaron los elementos de estructura secundaria extremos de la cadena y los giros 1 y 3 del receptor (tomada de Kirsch, T. Et al. 2000).

Mecanismo de acción de las BMP

Aunque las BMP son proteínas multifuncionales, las BMPs identificadas como osteogénicas, son básicamente las BMP-2 y BMP-7 que pueden inducir la formación de hueso y cartílago. Estas proteínas actúan como señales locales que difunden hasta la superficie celular, alterando el proceso de desarrollo. La BMP-2 actúa como un homodímero unido por un enlace de hidrógeno e induce la formación de hueso y cartílago. Juega un papel crítico en la diferenciación de osteoblastos.

Desde el punto de vista físico y químico, las BMPs se pueden considerar como proteínas secretadas que actúan como ligandos y que se unen a receptores específicos en la superficie celular de diferentes tipos de células. Al interactuar con su receptor, causan efectos autocrinos y/o paracrinos, con lo que tienen un impacto importante en la organización celular y tisular.

Los receptores transduce la señal mediante un grupo de proteínas conocidas como SMADS (este término es una fusión de los términos MAD, del inglés “mothers against decapentaplegic”, una proteína identificada en *Drosophila*, cuya forma mutada reprime el gen decapentaplégico, importante en el desarrollo embrionario, y el término SMA, una proteína encontrada en *Caenorhabditis elegans*, que causa un pequeño tamaño corporal en el nematodo (small). Las SMADS son homólogas a ambos tipos de proteínas) (Attisano y Wrana, 1998; Whitman, 1998; Derynck, 1998). Estos Smads llevan la señal desde el citoplasma hasta el núcleo, por lo que se les considera factores de transcripción que regulan la expresión de algunos genes particulares (Whitman, 1998).

In vivo, las BMPs actúan esencialmente como factores de diferenciación, convirtiendo células mesenquimáticas en células formadoras de cartílago y hueso (Derynck, 1998; Massagué et al., 2005). Un estudio sugiere que tanto in vitro como in vivo la implantación de estas BMPs, estimulan células en tejidos blandos y duros (músculo, médula ósea y periosteum), que se convierten en células formadoras de hueso ó de cartílago (Pogue y Lyons, 2006).

La actividad de las BMPs está controlada en varios niveles. El factor de transcripción tejido-específico (conocido también como factor básico hélice-giro-hélice) y su secuencia de unión (conocida como “caja E”), juegan un papel crítico en el camino de expresión que siguen las BMPs. En el medio extracelular, proteínas solubles inhibitorias como las noguinas, cordinas y folistatinas, pueden afectar a algunas BMPs e inhibir su unión a los receptores de superficie celular. En el interior de la célula, la actividad de las BMPs se controla mediante la combinación de la transducción de la señal y las proteínas inhibitorias Smad (Singhatanadgit y Olsen, 2006) (Figura 5.5).

Las BMPs pueden también regular la expresión de las proteínas inhibitorias Smad. Estas Smad están fosforiladas y se translocan en el núcleo, donde regulan la transcripción de genes diana, junto con otros factores de transcripción que incluyen PEBP2alphaA/Cbfa1 (Tou, L., Quibria, N. y Alexander, J.M. 2003).

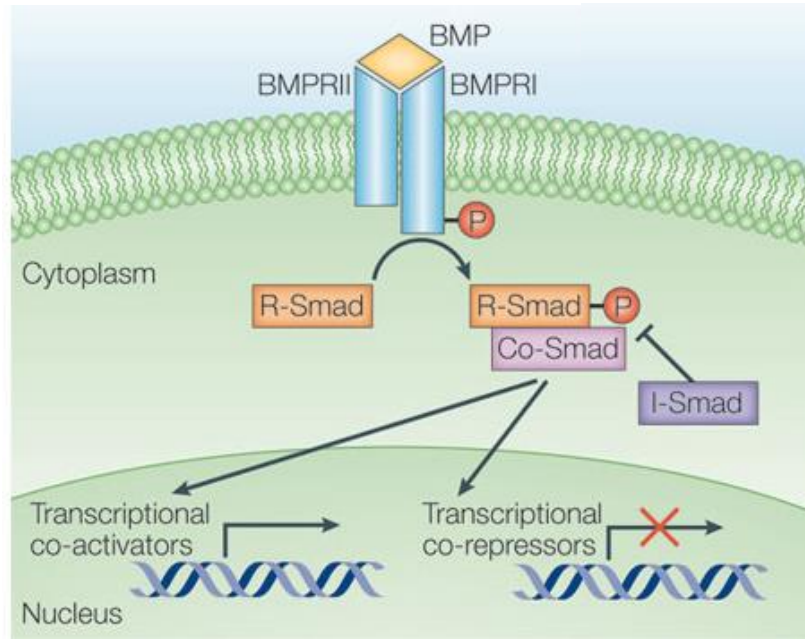


Figura 5.8. Mecanismo de acción de las BMP-2.

Otras funciones de las proteínas de la familia de las TGF- β

Las proteínas de la superfamilia de las TGF- β cumplen funciones muy variadas en el desarrollo de los organismos. De ellas, las usadas en el presente estudio para la reconstrucción de las proteínas ancestrales, es decir, las de las familias de proteínas parecidas a las BMPs, abarcan 19 grupos, que van desde las BMP-2 a las BMP-10, y de las BMP-12 a las BMP-15 y las GDF-1, 3, 5, 6, 7, 9 y Nodal, forman un conjunto de proteínas participantes en señalización de múltiples procesos celulares, funciones diferentes en diferentes tipos celulares, y un enorme potencial terapéutico, que se menciona brevemente al final de este trabajo. En la siguiente tabla, se presentan algunas de las funciones identificadas para las proteínas usadas en este trabajo, para la reconstrucción de proteínas ancestrales, y de las que como potencial perspectiva futura, este trabajo puede derivar en la reconstrucción de otras proteínas de la misma familia.

Tabla 6.1. Proteínas de la familia de los TGF- β y algunas de sus funciones.

Proteínas de la familia de los TGF- β	Funciones identificadas
BMP2	Producción de hueso y cartílago, usada también en implantes dentales y cirugía maxilofacial
BMP3	Regula negativamente la densidad ósea
BMP3B / GDF10	Involucrada en la formación del cráneo y en morfogénesis del esqueleto. Diferenciación del embrión en humanos. Diferenciación del tejido ectodérmico. Juega un papel importante en la formación de hueso endocondrial. Está involucrada en el desarrollo muscular y mineralización ósea.
BMP4	Implicada en el desarrollo de algunos cánceres. También en el desarrollo del nervio óptico.
BMP5	Induce a todos los marcadores osteogénicos en células madre mesenquimáticas.
BMP6	Transformación de células mesenquimáticas en hueso y cartílago. Induce diferenciación a osteoblastos.
BMP7	Participa en la osteogénesis epitelial. También en la homeostasis ósea.
BMP8A	Homóloga a BMP7 y BMP5. Parece estar implicada en la osteogénesis.
BMP8B	Participa en inducir y mantener la habilidad de las neuronas colinérgicas del prosencéfalo para responder a la acetilcolina. Estas neuronas son importantes para el aprendizaje, la memoria y la atención. También parece estar involucrada en la regulación del metabolismo del hierro.
BMP9 / GDF2	Involucrada en la tabicación del corazón.
BMP10	Regula los genes Hox, participantes en el desarrollo anteroposterior y el cordón espinal. Regula el número de receptores olfatorios y de células retinales ganglionicas en el desarrollo de la retina.
BMP11 / GDF11	Participa en el desarrollo nervioso del embrión en crecimiento. Induce la formación de neuronas sensoriales en el cordón de la espina dorsal.
BMP12 / GDF7	Crecimiento y diferenciación de embriones en desarrollo. También participa en el desarrollo de la epidermis y las articulaciones óseas.
BMP13 / GDF6	Bloquea el efecto de la Nogina, que controla el desarrollo neuronal.
BMP14 / GDF5	Participa en el desarrollo esquelético.
BMP15	Participa en el desarrollo y maduración de los ovocitos y folículos. Sólo se expresa en los ovarios.
GDF1	Participa en el desarrollo de los patrones izquierda-derecha y la inducción del mesodermo en el desarrollo embrionario.
GDF3 / Vgr2	Potencia la actividad de otros factores de crecimiento. Inhibe a algunas BMPs. Participa en la regulación del desarrollo de las células madre embrionarias. También en la formación del endodermo y mesodermo en los estados pre-gástrula.
GDF9	Afecta la formación y desarrollo de los ovocitos. Se requiere para la génesis de folículos en los ovarios.
Nodal	Esencial para la formación del mesodermo y la formación de estructuras del eje izquierdo-derecho en el desarrollo embrionario temprano.

6.3. OBJETIVOS

6.3.1. OBJETIVO GENERAL

Mediante herramientas de la Biología Sintética Evolutiva, reconstruir secuencias ancestrales de proteínas de la superfamilia de los Factores de Crecimiento Transformantes tipo Beta, para la obtención de secuencias de ancestros de Proteínas Morfogenéticas de Hueso tipo 2, para su evaluación en el laboratorio.

6.3.2. OBJETIVOS ESPECÍFICOS

Seleccionar secuencias actuales de proteínas de la superfamilia de los Factores de Crecimiento Transformantes tipo Beta, con base en la similaridad de secuencia, organismo de procedencia, disponibilidad de secuencia completa, y valores de confianza como el valor E.

Realizar alineamientos de las secuencias de las proteínas seleccionadas.

Poner a punto la metodología computacional de la Biología Sintética Evolutiva para la reconstrucción de proteínas ancestrales.

Obtener secuencias ancestrales con altos niveles de probabilidad a posteriori, para su resurrección y evaluación en el laboratorio.

6.4. MATERIALES Y MÉTODOS

Selección de la secuencia de inicio (secuencia “query”)

Usamos la secuencia de residuos de aminoácidos de la preproteína BMP-2 humana (código de referencia NCBI: NG_023233.1, número de acceso YR9FBSMP011), como punto de partida para hacer una búsqueda de BLAST de otros miembros de la familia de los TGF- β .

Búsqueda de secuencias en la base de datos del NCBI

Usamos la opción de búsqueda en genomas completos del Centro Nacional de Información Biotecnológica de los Estados Unidos, que es tal vez una de las fuentes más completas de recursos bioinformáticos para la investigación de genes y proteínas (<http://www.ncbi.nlm.nih.gov/genome>). Las condiciones de búsqueda incluyeron un umbral de 0.0001 que corresponde a la probabilidad de que dos secuencias sean similares sólo por azar, y un rango de identidad de 20 al 100%.

Realizamos la búsqueda en organismos cuyos genomas han sido completamente secuenciados, distribuidos así: 12 genomas de mamíferos : humano (*Homo sapiens*), orangután (*Pongo abelii*), mono Rhesus (*Macaca mulatta*), mono Tití (*Callithrix jacchus*), conejo (*Oryctolagus cuniculus*), rata (*Rattus norvegicus*), oso panda (*Ailuropoda melanoleuca*), perro doméstico (*Canis lupus familiaris*), Caballo (*Equus caballus*), toro (*Bos Taurus*), rata marsupial (*Monodelphis domestica*), y ornitorrinco (*Ornithorhynchus anatinus*); 3 aves: Gallo doméstico (*Gallus gallus*), Pavo real (*Meleagris gallopavo*), y pinzón cebra (*Taeniopygia guttata*); un reptil: el lagarto anolés (*Anolis carolinensis*); un anfibio: el Sapo tropical (*Xenopus* (*Silurana*) *tropicalis*); el pez cebra (*Danio rerio*), y cuatro invertebrados: el gusano de la bellota (*Saccoglossus kowalevskii*), el erizo púrpura de mar (*Strongylocentrotus purpuratus*), el nematodo *C.elegans* (*Caenorhabditis elegans*), y la mosca de la fruta (*Drosophila melanogaster*).

Alineamiento de las secuencias seleccionadas

Para el alineamiento, usamos el programa MUSCLE (Edgar, 2004a y b), con la matriz Gonnet250, considerar una de las mejores matrices evolutivas para proteínas. Hicimos los alineamientos y la edición de los mismos, en el programa MEGA (Tamura et al., 2007; Tamura et al., 2011). Una vez editado el alineamiento pasamos a construir los árboles filogenéticos.

Árboles filogenéticos de las secuencias seleccionadas

Con base en el alineamiento, construimos múltiples árboles filogenéticos usando el programa MrBayes, basado en Inferencia Bayesiana. Este programa usa métodos de cadenas de Markov y muestreo de Montecarlo para estimar la distribución de los parámetros del modelo (Huelsenbeck et al., 2001. Huelsenbeck y Ronquist, 2001). Se realizaron decenas de árboles

filogenéticos para su evaluación a la luz de la información de la literatura y las relaciones evolutivas del “Árbol de la vida” (“Timetree”, <http://www.timetree.org/>; Hedges et al., 2006); sin embargo, no fue posible resolver dos politomías en el árbol, por lo que optamos por adaptar el árbol que se había obtenido, al árbol filogenético de los TGF- β , publicado recientemente en Nature, por Shi et al. (2011).

Reconstrucción de las secuencias ancestrales de los TGF- β

Para la reconstrucción alimentamos el programa PAML, con los alineamientos y el árbol filogenético obtenido; PAML usa el modelo de Máxima Verosimilitud, para reconstruir las secuencias ancestrales más probables, mediante el modelo de Jones (Yang, 1997; Yang, 2007). Posteriormente, analizamos las probabilidades a posteriori, y las mutaciones obtenidas, para tomar las decisiones sobre qué secuencias reconstruidas, se resucitarán.

6.5. RESULTADOS Y DISCUSIÓN

La búsqueda tipo BLASTP que hicimos en los genomas completos de los organismos mencionados, arrojó secuencias de proteínas de las siguientes subfamilias de la superfamilia de TGF- β : BMP-2, BMP-3, BMP-3B (GDF-10), BMP-4, BMP-5, BMP-6, BMP-7, BMP-8A, BMP-8B, BMP-9 (GDF-2), BMP-10, BMP-12 (GDF-7), BMP-13 (GDF-6), BMP-14 (GDF-5), BMP-15, GDF-1, GDF-3, GDF-9, y Nodal (GDF = Factor de crecimiento y diferenciación, por su sigla del inglés); todas estas proteínas pertenecen a la familia de los Factores de Crecimiento Transformantes tipo β , similares a Proteínas Morfogenéticas de Hueso. Para la escogencia de estas subfamilias, seguimos los criterios de clasificación de los TGF- β , más recientes y aparentemente más robustos (Shi et al., 2011). Hicimos una ronda adicional de selección de secuencias, buscando eliminar duplicados, proteínas anotadas incorrectamente, proteínas secuenciadas parcialmente ó con aparentes errores de secuenciación, e incluimos secuencias de proteínas de la superfamilia y de los organismos seleccionados, recién reportadas, pero que no estaban reportadas aún en los genomas completos de los mismos, con lo que se obtuvieron las secuencias que aparecen relacionadas en la tabla 2.

Tabla 5.2. Secuencias de proteínas de la familia de lasTGF-B seleccionadas, tanto por tipo de subfamilia proteica, como por organismo. El alineamiento se realizó con estas 323 secuencias.

Especies	BMP 2	BMP 3	GDF10 BMP3B	BMP 4	BMP 5	BMP 6	BMP 7	BMP 8A	BMP 8B	BMP 9	GDF2 10	BMP 12	GDF7 BMP13	GDF6 BMP14	GDF5 15	BMP 1	GDF 3	GDF 9	Nodal	Tot
<i>Homo sapiens</i>	1	1	1	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	21
<i>Pongo abelii</i>	1	1	1	1	1		1		2	1	1	1	1	1	1	1	1	3	1	19
<i>Macaca mulatta</i>	1	1	1	4	1	1	1	1		1	1	1	1	1	1	1	1	1	1	20
<i>Callithrix jacchus</i>	1	1	1	1	1	1	1		1	1	1	1	1	1	1	1	1	1	1	18
<i>Oryctolagus cuniculus</i>	1	1	1	1	1		1			1	1	1	1	1	1	1		1	1	15
<i>Rattus norvegicus</i>	1	1		1	1	1	2	2			1			1	1	1	2	1	1	19
<i>Ailuropoda melanoleuca</i>	1	1	1	1	1		1			1	1				1	1		2	1	14
<i>Canis lupus familiaris</i>	1		1		1	1	2		1	1	1				1			1	2	14
<i>Equus caballus</i>	1		1	1	1	1	1			1	1			1	1	1		1	1	14
<i>Bos taurus</i>	1	1	1	1	2	2	2	3		1	2	1	1	1	1	1	2	1	1	25
<i>Monodelphis domestica</i>	1	1	1	1	1	1	1			1	1	1	1	1	1	2	1	1	1	18
<i>Ornithorhynchus anatinus</i>	1	1	1				2	1		1	1			1		1		1		11
<i>Gallus gallus</i>	1	1	1	1	1	1	1			1	1				1	1		1	1	14
<i>Meleagris gallopavo</i>	1	1	1		1		1	1		1	1				1	1		1	1	13
<i>Taeniopygia guttata</i>				1	1		1				1				1	1		1	1	10
<i>Anolis carolinensis</i>	1	1												2	1			1	1	7
<i>Xenopus (Silurana) tropicalis</i>	2	1	1	1	1	1	2		1	1	1	1	1	1	1	1	1	1	1	27
<i>Danio rerio</i>	1	1	2	1	2	1	2	1		1	2			3	1	1		1	2	22
<i>Saccoglossus kowalevskii</i>	1	1		1	2						1								3	9
<i>Strongylocentrotus purpuratus</i>	2	2			1		1	1											2	9
<i>Caenorhabditis elegans</i>	2																			2
<i>Drosophila melanogaster</i>				1			1													2
Totales	23	19	16	21	21	12	25	8	9	15	20	8	16	17	17	9	17	19	31	323

Para realizar el alineamiento añadimos una secuencia de la misma superfamilia de los TGF- β , pero que es conocido que pertenece a otra subfamilia, más lejanamente relacionada desde el punto de vista evolutivo. La razón de incluir esta secuencia, fue la de proveer una secuencia “outgroup”, que sirviera para enraizar el árbol filogenético obtenido. La secuencia que escogimos para esto, fue la de la proteína GDNF de ratón (GDNF : factor neurotrófico derivado de línea de células gliales, del inglés “Glial cell line-derived neurotrophic factor”). En la Figura 5.5 se presenta una parte del alineamiento obtenido con el programa Muscle.

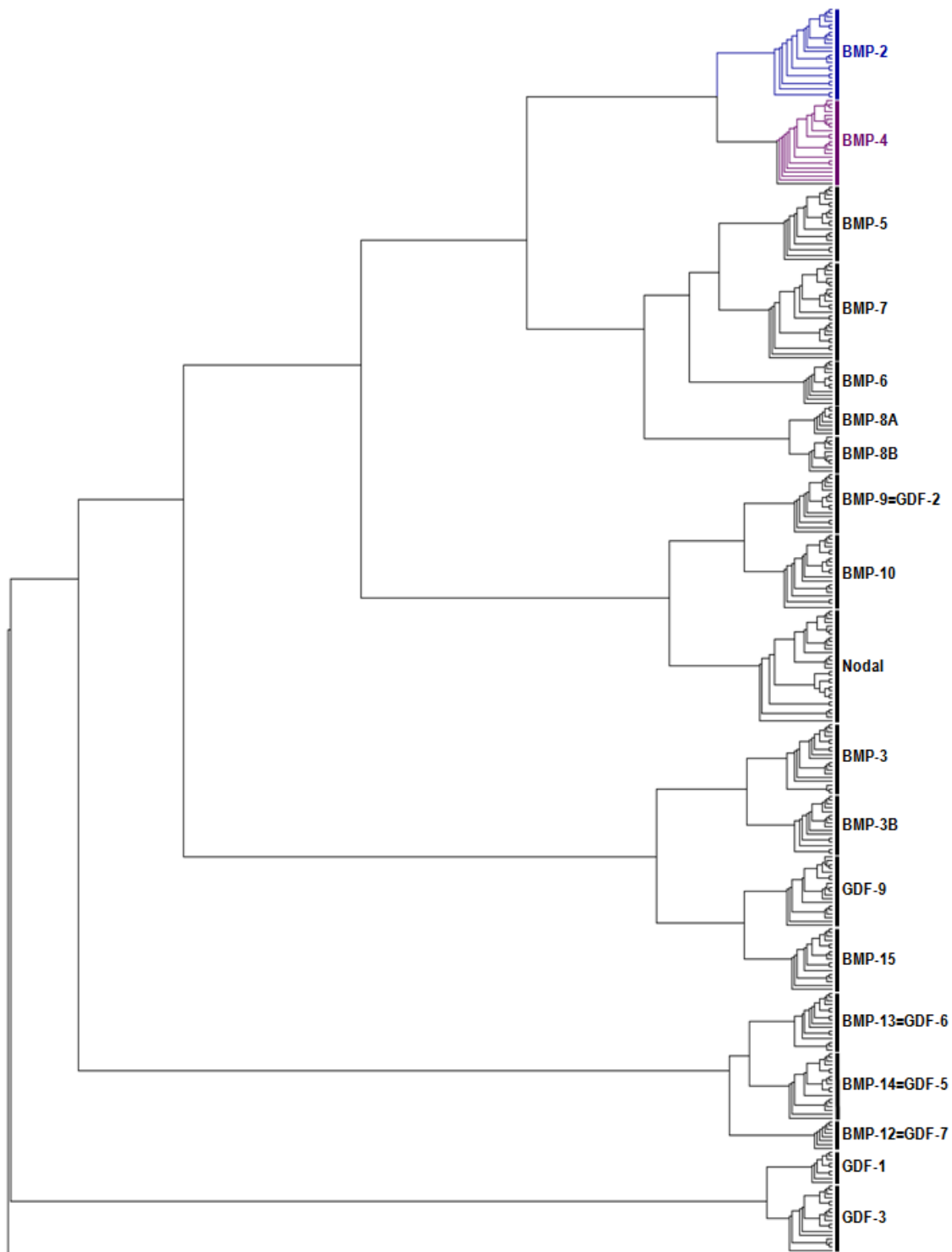


Figura 6.10. Árbol filogenético de la familia de proteínas parecidas a BMPs, de la superfamilia de las TGF- β .

Este árbol filogenético tiene algunas particularidades: las ramas terminales corresponden a un tipo de proteínas; una subfamilia (ejemplo BMP-2, BMP-4, etc). Además de ello, como puede observarse en la Figura 5.8, cada rama terminal, incluye secuencias de los organismos seleccionados; Es decir, el árbol filogenético que hemos obtenido presenta relaciones evolutivas

entre especies animales en sus ramas terminales (por ejemplo *Homo sapiens* y *Drosophila melanogaster*), y relaciones evolutivas entre familias de proteínas en los nodos más antiguos (por ejemplo las BMP-2 y las BMP-7), con lo que podemos inferir tanto eventos relacionados con los procesos de especiación (proteínas ortólogas), como eventos relacionados con la duplicación de genes (proteínas parálogas). Así, son ortólogas, la proteína BMP-2 humana, con la proteína BMP-2 de orangután, mientras que son ortólogas, las proteínas BMP-2 y las BMP-4.

Debido a lo anterior, en las ramas terminales podemos estimar distancias evolutivas entre especies (por ejemplo, tenemos la distancia entre *Homo sapiens* y *Drosophila melanogaster*, cuyo ancestro común está a 782.7 millones de años en el pasado). Podemos estimar también la distancia evolutiva entre proteínas parálogas. Sin embargo, en el presente reporte, no presentamos la distancia evolutiva entre estas proteínas.

Todos estos nodos reconstruidos son más antiguos que la mayor distancia evolutiva entre las especies seleccionadas para esta reconstrucción; es decir, son más antiguos que 782.7 millones de años (ilustraciones 5.7 y 5.8). De los nodos reconstruidos computacionalmente, presentamos aquí tres nodos seleccionados para su resurrección y estudio en el laboratorio.

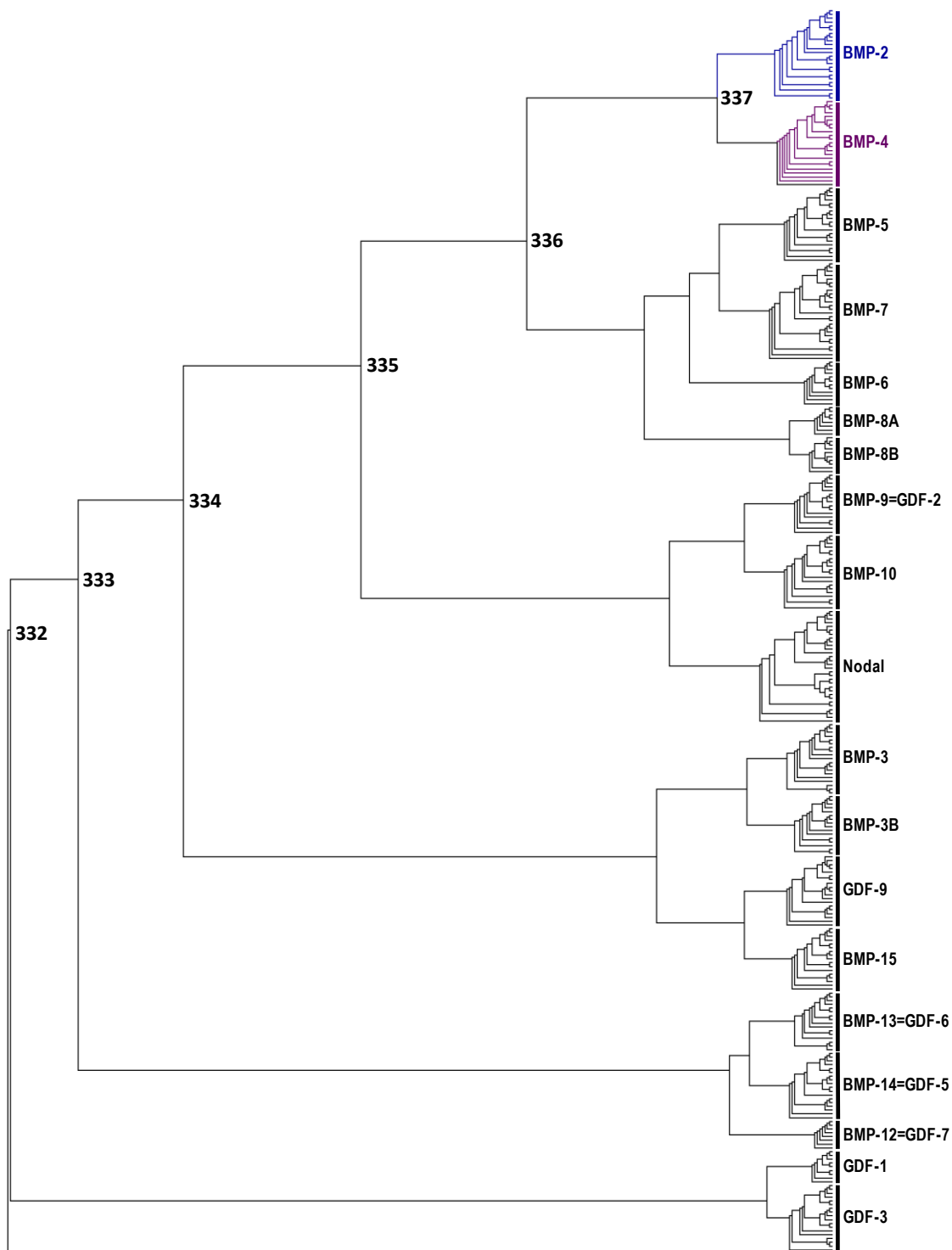


Figura 6.11. Arbol filogenético mostrando algunos de los nodos ancestrales, de los cuales se reconstruyo el nodo 336 para resucitar en el laboratorio.

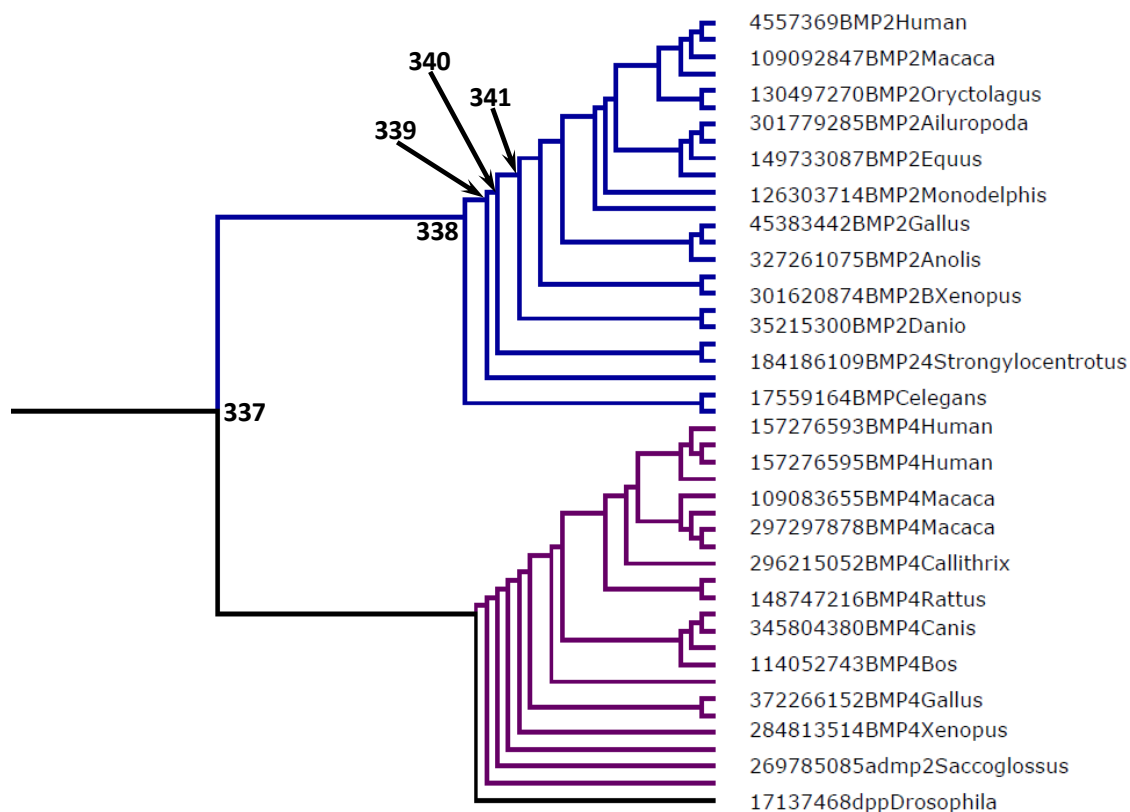


Figura 6.12. Detalle del árbol filogenético de las TGF- β , mostrando las ramas de las BMP-2 y las BMP-4. Se muestran los nodos ancestrales más cercanos a las proteínas BMP-2. Se reconstruyeron los nodos 340 y 341.

En las figuras 5.11 y 5.12 se presentan los nodos ancestrales reconstruidos (los nodos 336, 340 y 341), que corresponden a las siguientes interpretaciones evolutivas:

Nodo número 336: Ancestro común entre el grupo de subfamilias de proteínas BMP-2 / BMP-4, y el grupo de subfamilias BPM-5 / BMP-6 / BMP-7 / BMP-8A / BMP-8B.

Nodo número 340: Ancestro común entre las proteínas BMP-2 de vertebrados y las de invertebrados.

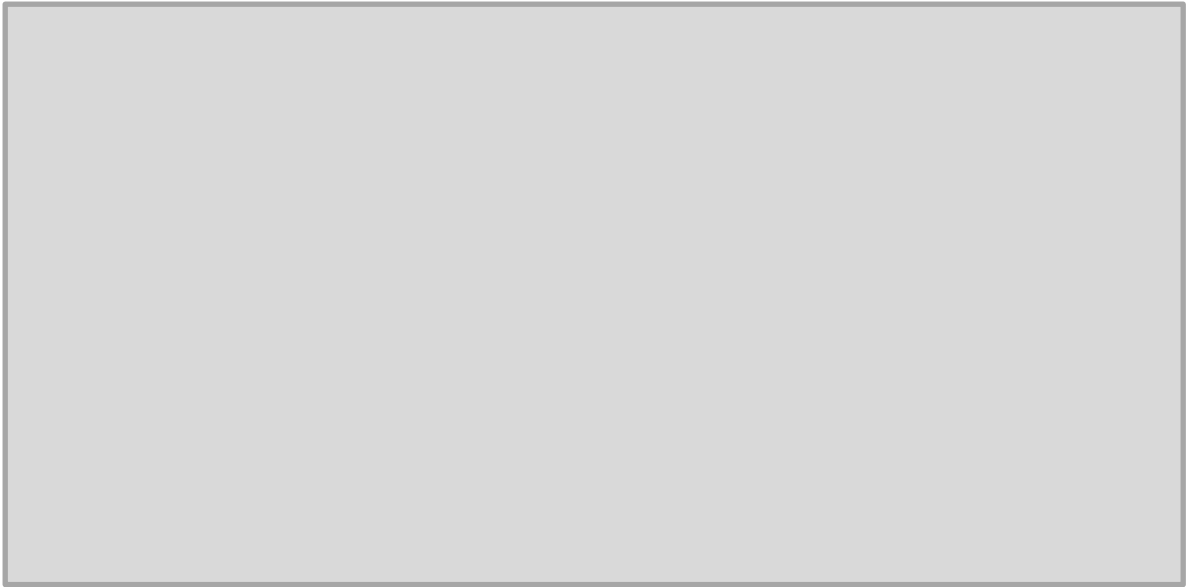
Nodo número 341: Ancestro común entre las proteínas de mamíferos / aves / reptiles, y las proteínas de peces e invertebrados.

Los nodos 336 y 337 presentaron las mismas secuencias, con 7 residuos diferentes de la proteína BMP-2 humana. Los nodos 342, 343 y 344, también muestran secuencias iguales, con sólo tres mutaciones con respecto a la BMP-2 humana; y los nodos 345 hasta las proteínas BMP-2 actuales, presentaron la misma secuencia que la proteína BPM-2 humana.

Las reconstrucciones seleccionadas para estos tres nodos, fueron las realizadas con los modelos WAG2, mediante reconstrucción marginal.

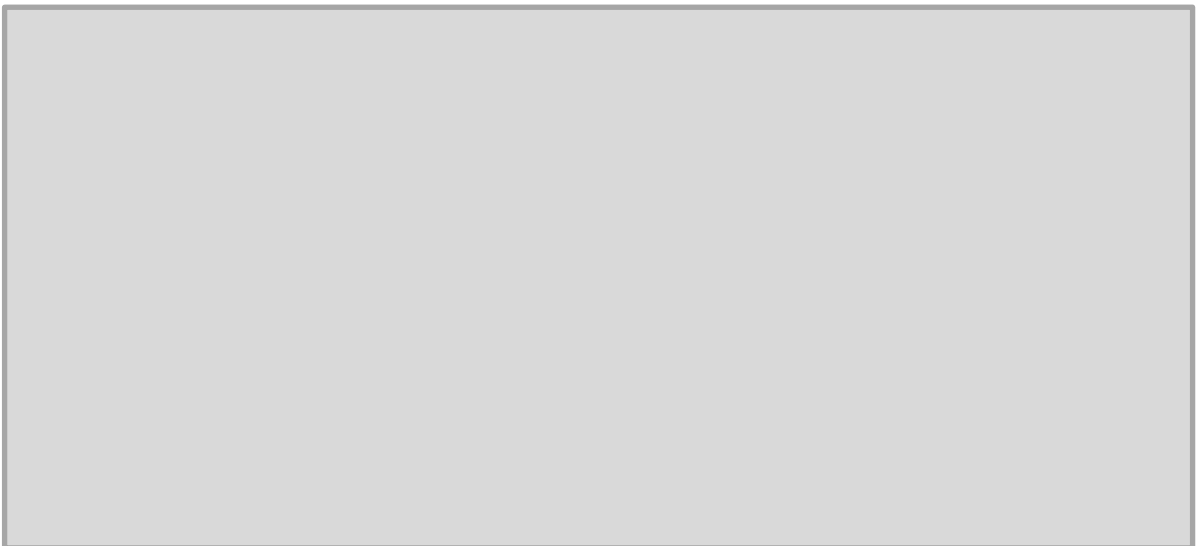
NOTA: debido a que las proteínas rediseñadas obtenidas en este trabajo, son un producto con el que aplicará a una patente, la oficina de transferencia de Tecnología de la Universidad de Granada, ha solicitado expresamente la omisión de esa información en este documento, con el fin de evitar la vulneración de la aplicación a la patente. Por ello, se ha ocultado en las siguientes tablas y figuras, que presentan las secuencias de las proteínas reconstruidas.

Tabla 2. Nodos reconstruidos, con sus mutaciones y porcentajes de similitud y disimilaridad con respecto a la BMP-2 humana.



Como se nota en la tabla 2. Los nodos más cercanos a las proteínas actuales, presentan un menor número de mutaciones y por tanto más similitud con las BMPs actuales, que los nodos más antiguos.

Las secuencias de las proteínas ancestrales obtenidas, son



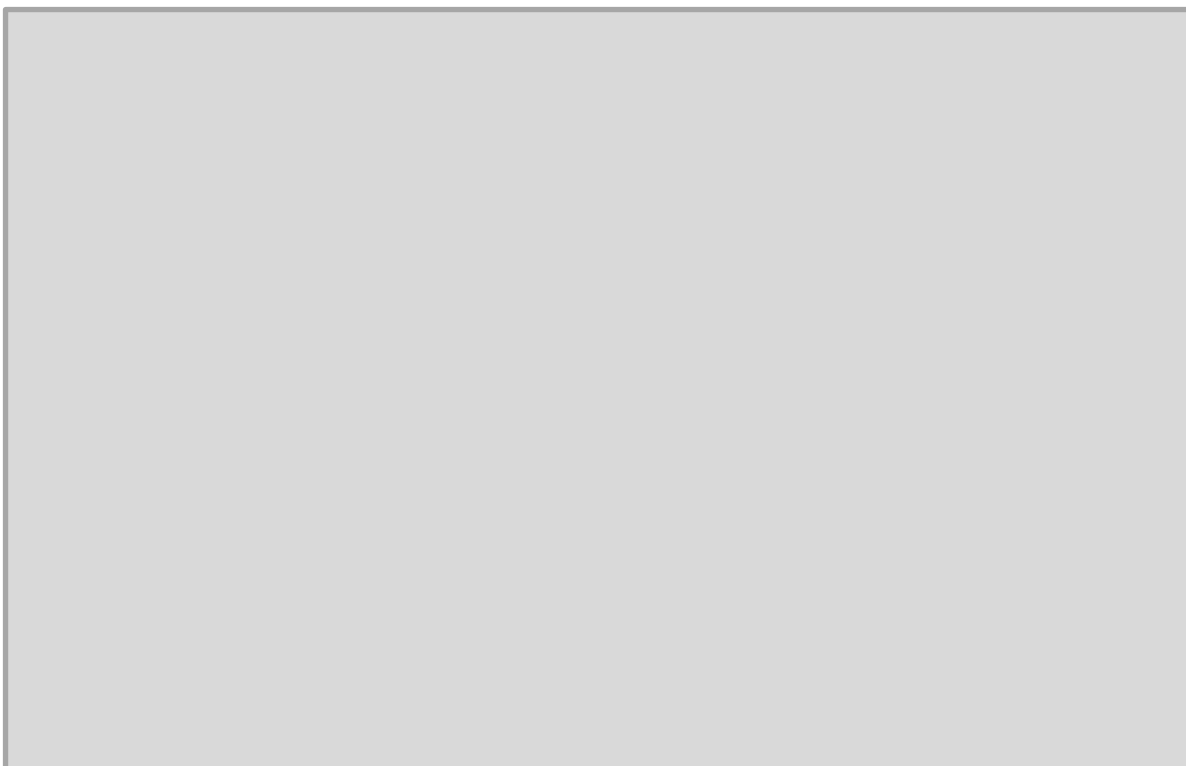


Figura 5.13. Mutaciones de los nodos ancestrales reconstruidos, con relación a la BMP-2 humana. La primera línea corresponde a las posiciones de los residuos en la proteína. La segunda línea es la secuencia de la BMP-2 humana. Las tercera, cuarta y quinta líneas, corresponden a las mutaciones en las secuencias de las proteínas ancestrales de los nodos 336, 340 y 341.

Note que dos de los residuos importantes para la unión al receptor, Leu51 y Asp53 (marcados con un recuadro rojo en la Figura 6.9), no están modificados en el nodo ancestral 41, mientras que en los nodos más antiguos (36 y 40), ocurre una mutación por un aminoácido equivalente, conservando las propiedades fisicoquímicas (Glu por Asp), pero no mutación en la Leu51.

También es importante resaltar que las zonas grises en la Figura 6.9, corresponden a zonas de interacción con el receptor, y que con la excepción del cambio de Asn por Tyr en la posición 96, la proteína ancestral del nodo 41 (es decir, el nodo reconstruido más cercano a las proteínas actuales). En cuanto a los residuos que establecen enlaces de hidrógeno con el receptor, las proteínas ancestrales no presentan mutaciones en las posiciones W28, L51, D53, N59, S69, e Y103. En las ancestrales de los nodos 36 y 40, se ha cambiado un Glu53 por un Asp, por lo que la mutación es conservativa y se puede mantener el enlace de hidrógeno. Con esto, al parecer no hay mutaciones en las zonas importantes de la proteína para la interacción con el receptor.

A medida que los nodos reconstruidos se alejan de las proteínas actuales, aparecen más mutaciones que pueden (*a priori*), afectar la unión del ligando (la BMP-2) con su receptor de membrana. Hasta el momento, se tienen construidos los plásmidos con los genes de las proteínas ancestrales, para su expresión tanto en *Escherichia coli*, como en células de ovario de hámster chino (células CHO), como modelo de mamífero.

6.6. Sobre la Reconstrucción Ancestral de β -Lactamasas

Se reconstruyeron enzimas ancestrales de enzimas β -Lactamasas clase A, correspondientes diferentes épocas de la historia del planeta, llegando hasta periodos precámbricos.

Se resucitaron Lactamasas correspondientes a los siguientes ancestros comunes: **ENCA** (Enterobacteria Common Ancestor, 600 millones de años atrás), **GPBCA** (Gamma Proteo Bacteria Common Ancestor, de 1500 millones de años de antigüedad), **GNCA** (Gram Negative Common Ancestor, 2000 millones de años atrás), y **PNCA** (Gram Positive/Negative Common Ancestor, de hace 3000 millones de años). Se usó una enzima β -Lactamasas clase A existente hoy en día, en *E. coli*, como secuencia semilla para la recuperación de las secuencias de las demás lactamasas de origen cromosómico, de hoy en día.

Las enzimas presentaron secuencias sustancialmente diferentes a las enzimas de hoy: **ENCA** (55 residuos diferentes), **GPBCA** (111 diferentes), **GNCA** (122 diferentes), y **PNCA** (124 diferentes).

Todas las proteínas ancestrales presentaron una T_m mayor que las lactamasas de hoy, llegando hasta una estabilidad térmica de 35°C mayor, lo que es coherente con la temperatura del planeta de hace miles de millones de años, según el registro geológico y paleoclimático. El lograr termoestabilidades mayores que las proteínas de hoy, es algo deseable en la ingeniería de proteínas, dado que se logran altas temperaturas de actividad en procesos industriales, y se elimina la necesidad de mantener una cadena de frío de una enzima, por ejemplo.

Las enzimas ancestrales reconstruidas presentaron también una mayor promiscuidad de sustrato, que también es coherente con la teoría evolutiva, en que la especialización se logra con el paso de mucho tiempo de respuesta a condiciones cambiantes que ejercen presión selectiva sobre los genes de las enzimas.

En esta sección, se presenta el artículo “Hyperstability and substrate promiscuity in laboratory resurrections of Precambrian β -Lactamasas”, publicado en el Journal of American Chemical Society en febrero de 2013, que incluye la metodología, usada tanto en la reconstrucción computacional, como el trabajo experimental que llevó a la resurrección y caracterización de las enzimas.

Mi participación en la reconstrucción de Lactamasas, correspondió a la reconstrucción computacional de las mismas, clonación de los genes en los vectores adecuados, expresión de las proteínas, y realización de algunas de las pruebas de microcalorimetría para verificar la estabilidad térmica de las enzimas ancestrales.

A continuación, el artículo publicado que pretende mostrar lo que se presentaría en esta sección del capítulo 6. En el capítulo 7, se presentan los productos derivados de este trabajo.

6.6. CONCLUSIONES CAPÍTULO 6 - REDISEÑO DE BMPs y β -LACTAMASAS MEDIANTE RECONSTRUCCIÓN ANCESTRAL

Las proteínas hipotéticamente ancestrales, resucitadas en el laboratorio, presentaron la actividad que se esperaba, que es la misma de las proteínas “semilla” usadas para la selección de secuencias proteicas iniciales.

Las proteínas hipotéticamente ancestrales presentan una estabilidad térmica magnificada, y temperaturas de fusión (T_m), hasta 35 grados centígrados más alta.

La termoestabilidad magnificada de las proteínas ancestrales correspondientes a períodos precámbricos, es coherente con la evidencia geológica que dice que el planeta estaba a mayores temperaturas que hoy en día.

Las proteínas ancestrales presentan mayor promiscuidad en su actividad que las proteínas de hoy, lo que es coherente con la teoría evolutiva.

La reconstrucción de secuencias ancestrales se constituye en una herramienta poderosa para la ingeniería de proteínas, sobre todo cuando se requieren enzimas más termoestables ó se requiere promiscuidad de sustrato.

La Reconstrucción de Secuencias Ancestrales permite hacer inferencias evolutivas, sobre genes y proteínas, y eventualmente usar esa información para el rediseño de proteínas, genes y genomas de hoy en día.

SIGUIENTES PASOS Y PERSPECTIVAS FUTURAS

El paso lógico siguiente, es evaluar en el laboratorio las proteínas ancestrales reconstruidas, al igual que evaluar su actividad in vitro e in vivo. Para ello, ya se han construido los plásmidos con los genes de estas proteínas. Se expresarán tanto en *E. coli*, como en células de mamífero (células CHO, de Ovario de Hamster Chino), por lo que se han construido los genes tanto en un plásmido de expresión para *E. coli*, como para la expresión en células de mamífero.

Aunque en el presente trabajo se trata la reconstrucción de sólo tres nodos ancestrales, hemos reconstruido en realidad todos los nodos ancestrales del árbol filogenético de las ilustraciones 7 y 8. Esto significa que podemos evaluar nodos más antiguos como candidatos a modificaciones en las proteínas BMP-2.

Lo anterior también significa que podemos reconstruir nodos de otras proteínas de interés médico, como las BMP-7 que se usan para estimular el crecimiento de células renales, y de esta manera aumentar el éxito del trasplante de riñón. Esta proteína también participa en la recuperación de fibras musculares cardíacas como tratamiento posterior al infarto de miocardio, y en la rehabilitación motriz de algunas áreas dañadas del cerebro. Hay otras proteínas de la superfamilia de las TGF- β , que tienen otras potenciales aplicaciones terapéuticas, con lo que las posibilidades que abre la reconstrucción ancestral de estas proteínas son muchas y de variadas aplicaciones potenciales.

BIBLIOGRAFÍA - REDISEÑO DE BMPs Y β -LACTAMASAS MEDIANTE RECONSTRUCCIÓN ANCESTRAL

Álvarez-Nebreda ML, Jiménez AB, Rodríguez P, Serra JA. (2008). **Epidemiology of hip fracture in the elderly in Spain**. Bone. 42:278-85.

Adey, N.B., Tollefsbol, T.O., Sparks, A.B., Edgell, M.H. y Hutchison, C. A. (1994). **Molecular resurrection of an extinct ancestral promoter for mouse L1**. Proc. Natl. Acad. Sci. U.S.A. 91:1569–1573.

Attisano L, Wrana JL (1998). **Mads and Smads in TGF beta signalling**. Current Opinions in Cell Biology. 10 (2): 188–94.

Benner, S.A. y Allemann, R.K. (1989). **The return of pancreatic ribonucleases**. Trends in Biochemical Sciences. 14: 396–397.

Bleuming SA, He XC, Kodach LL, Hardwick JC, Koopman FA, Ten Kate FJ, van Deventer SJ, Hommes DW, Peppelenbosch MP, Offerhaus GJ, Li L, van den Brink GR. (2007). **Bone morphogenetic protein signaling suppresses tumorigenesis at gastric epithelial transition zones in mice**. Cancer Research 67 (17): 8149–55.

Burt DW, Law AS (1994). **Evolution of the transforming growth factor-beta superfamily**. Prog. Growth Factor Research. 5 (1): 99–118.

Burt, D.W. (1992). **Evolutionary grouping of the Transforming Growth Factor- β superfamily**. Biochemical and Biophysical Research Communications. 184 (2): 590-595.

Chang, B.S. (2003). **Ancestral Gene Reconstruction and Synthesis of Ancient Rhodopsins in the Laboratory**. Integrative and Comparative Biology. 43 (4): 500-507.

Chang, B.S., Jönsson, K., Kazmi, M.A., Donoghue, M.J. y Sakmar, T.P. (2002). **Recreating a functional ancestral archosaur visual pigment**. Molecular Biology and Evolution, 19(9): 1483–1489.

Cochran, D.L., Jones, A.A., Lilly, L.C., Fiorellini, J.P. y Howell, H. (2000). **Evaluation of recombinant human bone morphogenetic protein-2 in oral applications including the use of endosseous implants: 3-year results of a pilot study in humans**. Journal of Periodontology. 71(8):1241-57.

Constam DB & Robertson EJ (1999). **Regulation of bone morphogenetic protein activity by prodomains and proprotein convertases**. Journal of Cell Biology, 144: 139-149.

Dean, A.M. y Golding, G.B. (1997). **Protein engineering reveals ancient adaptive replacements in isocitrate dehydrogenase**. Proc. Natl. Acad. Sci. USA, 94: 3104–3109.

Derynck R, Zhang Y, Feng XH (1998). **Smads: transcriptional activators of TGF-beta responses**. Cell 95 (6): 737–40.

Ebara S, Nakayama K. (2002). **Mechanism for the action of bone morphogenetic proteins and regulation of their activity**. Spine (Phila Pa 1976). 15;27(16 Suppl 1):S10-5.

Eck RV, Dayhoff MO (1966) **Atlas of protein sequence and structure**. National Biomedical Research Foundation, Silver Spring, MD. 161-202.

Edgar, R.C. (2004a). **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. Nucleic Acids Research. 32(5):1792-1797.

Edgar, R.C. (2004b). **MUSCLE: a multiple sequence alignment method which reduced time and space complexity**. BMC Bioinformatics. 5:113-132.

Einhorn T A, Majeska R J, y Mohaideen A. (2003). **A single injection of recombinant human bone morphogenetic protein-2 accelerates fracture repair**. J. Bone Joint Surg.Am. 85-A (1):425–435.

Fitch. (1971). **Toward defining the course of evolution: minimum change for a specific tree topology**. Systematic Zoology, 20: 406-416.

Friedlaender G E, Perry C R, Cole J D et al. (2001). **Osteogenic protein-1 (bone morphogenetic protein-7) in the treatment of tibial non-unions**. J. Bone Joint Surg.Am. 83-A suppl. 1:S pp. 151–158.

Govender S, Csimma C, Genant H K et al. (2002). **Recombinant human bone morphogenetic protein-2 for treatment of open tibial fractures: A prospective, controlled, randomized study of four hundred and fifty patients**. J. Bone Joint Surg.Am. 84-A(2):123–2,134.

Greenwald, A.S., Boden, S.D., Barrack, R.L., Bostrom, M., Goldberg, V.M., Yaszemski, M.J. y Heim, C.S. (2010). **The evolving role of graft-bone substitutes**. Proceedings of the American Academy of Orthopaedic Surgeons, Annual Meeting, 2010. 98-013.

Griffith DL, Keck PC, Sampath TK et al. (1996). **Three-dimensional structure of recombinant human osteogenic protein 1: structural paradigm for the transforming growth factor beta superfamily**. Proceedings of the National Academy of Sciences, USA, 93: 878-883.

Hak, D.J. (2005). **Current issues in orthopaedic trauma surgery**. US musculoskeletal review. 1:16-17

Hedges, S.B., Dudley, J. y Kumar, S. (2006). **TimeTree: a public knowledge-base of divergence times among organisms**. Bioinformatics. 22:2971-2972.

Herrera A, Martínez AA, Ferrandez L, Gil E, Moreno A. (2006). **Epidemiology of osteoporotic hip fractures in Spain**. International Orthopaedics. 30:11-4

Herpin A, Lelong C, Favrel P (2004). **Transforming growth factor-beta-related proteins: an ancestral and widespread superfamily of cytokines in metazoans.** Developmental and Comparative Immunology. 28 (5): 461–85.

Howell TH, Fiorellini J, Jones A et al. (1997). **A feasibility study evaluating RhBMP-2/absorbable collagen sponge device for local alveolar ridge preservation or augmentation.** International Journal of Periodontics and Restorative Dentistry, 17: 124-139.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen y J. P. Bollback. (2001). **Bayesian inference of phylogeny and its impact on evolutionary biology.** Science 294: 2310-2314.

Huelsenbeck, J. P. y F. Ronquist. (2001). **MRBAYES: Bayesian inference of phylogeny.** Bioinformatics 17:754-755.

Huelsenbeck, J. P., y J. P. Bollback. 2001. **Empirical and hierarchical Bayesian estimation of ancestral states.** Systematic Biology 50:351-366.

Jermann, T.M., Opitz, J.G., Stackhouse, J. y Benner, S. (1995). **Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily.** Nature 374, 57-59

Kalfas, I.J. (2001). **The principles of bone healing.** Neurosurgery focus. 10 (4): 1-4.

Kenley, R.A., Yim, K., Abrams, J., Ron, E., Turek, T., Marden, L.J. y Hollinger, J.O. (1993). **Biotechnology and Bone Graft Substitutes.** Pharmaceutical Research, 10: (10): 1393-1401

Kirk, O., T.V. Borchert, y C.C. Fuglsang. (2002). **Industrial enzyme applications.** Current opinions in biotechnology. 13(4): 345-51.

Kirsch T, Sebald W, Dreyer MK. (2000). **Crystal structure of the BMP-2-BRIA ectodomain complex.** Nature Structural Biology. 7(6):492-496.

Keller, S., Nickel, J, Zhang, J.L. Sebald, W. y Mueller, T.D. (2004). **Molecular recognition of BMP-2 and BMP receptor IA.** Nature Structural & Molecular Biology 11, 481 – 488.

Kodach LL, Wiercinska E, de Miranda NF, Bleuming SA, Musler AR, Peppelenbosch MP, Dekker E, van den Brink GR, van Noesel CJ, Morreau H, Hommes DW, Ten Dijke P, Offerhaus GJ, Hardwick JC. (2008). **The bone morphogenetic protein pathway is inactivated in the majority of sporadic colorectal cancers.** Gastroenterology 134 (5): 1332–41.

Koshi, J.M. y Goldstein, R.A. **Probabilistic Reconstruction of Ancestral Protein Sequences.** Journal of Molecular Evolution (1996) 42:313–320.

Lazar, G.A., et al. (2003). **Designing proteins for therapeutic applications.** Current opinion in structural biology. 13(4): p. 513-8.

Lacroix, P. (1945). **Recent investigations on the growth of bone.** Nature 156, 576-576

Levander, G. (1934). **On the formation of new bone in bone transplantation**. Acta Chir Scand 74: 425–426.

Levander, G. (1938). **A study of bone regeneration**. Surgery, Gynecology and Obstetrics. 67: 705–714.

Liberles, D. (2007). **Ancestral Sequence Reconstruction**. Oxford University Press. 272 páginas.

Luyten, F. P.; et al. (1989). **Purification and Partial Amino Acid Sequence of Osteogenin, a Protein Initiating Bone Differentiation**. Journal of Biological Chemistry 264 (23): 13377–13380.

Luyten, F.P., Cunningham, N.S., Ma, S., Muthukumaran, N., Hammonds, R.G., Nevins, W.V., Wood, W.I., y Reddi, A.H. (1989). **Purification and partial amino acid sequence of Osteogenin, a protein initiating bone differentiation**. Journal of Biological Chemistry. 264 (23): 13377-13380.

Malcolm, B.A., Wilson, K.P., Matthews, B.W., Kirsch, J.F. y Wilson, A.C. (1990). **Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing**. Nature, 345(6270):86–89.

Massagué J, Seoane J, Wotton D (2005). **Smad transcription factors**. Genes Dev. 19 (23): 2783–810.

McKay, W.F., Peckhan, S.M. y Badura, J.M. (2007). **A comprehensive clinical review of recombinant human bone morphogenetic protein-2 (INFUSE® bone graft) - Int Orthop**. December; 31(6): 729–734.

Milano F, van Baal JW, Buttar NS, Rygiel AM, de Kort F, DeMars CJ, Rosmolen WD, Bergman JJ, Van Marle J, Wang KK, Peppelenbosch MP, Krishnadath, KK. (2007). **Bone morphogenetic protein 4 expressed in esophagitis induces a columnar phenotype in esophageal squamous cells**. Gastroenterology 132 (7): 2412–21.

Moghadam HG et al., HG; Urist, MR; Sandor, GK; Clokie, CM (2001). **Successful mandibular reconstruction using a BMP bioimplant**. J Craniofac Surg. 12 (2): 119–127.

Murray-Rust, J., McDonald, N.Q., Blundell, T.L., Hosang, M., Oefner, C., Winkler, F., Bradshaw, R.A. (1993). **Topological similarities in TGF- β 2, PDGF-BB and NGF define a superfamily of polypeptide growth factors**. Structure. 1:153-159.

Nakashima M. y Reddi, AH. (2003). **The application of bone morphogenetic proteins to dental tissue engineering**. Nature Biotechnology. 21(9):1025-32.

Nandi, S.K., Roy, S., Mukherjee, P., Kundu, B., De, D.K. y Basu, D. (2001). **Orthopaedic applications of bone grafts and substitutes, a review**. Indian Journal of Medical Research. 132: 15-30.

Nogués Solán X, Guerri R, Solé E, Díez-Pérez A. (2010). **Impacto socioeconómico de la osteoporosis**. Revista de Osteoporosis y Metabolismo Mineral. 2010;2 (Supl 3): S8-S11

Pauling L. y Zuckerkandl E. (1963). **Chemical paleogenetics, molecular restoration studies of extinct forms of life**. Acta chemica Scandinavica 17 (89): 9–16

Perez-Jimenez, R., Inglés-Prieto, A., Sanchez-Romero, I., Alegre-Cebollada, J., Kosuri, P., Garcia-Mañes, S., Gaucher, E., Sanchez-Ruiz, J.M. y Fernandez, J.M. (2011). **Paleoenzymology at the Single-Molecule Level: Probing the Chemistry of Resurrected Enzymes with Force-Clamp Spectroscopy**. Nature Structural Biology. 18, 592–596.

Pogue, R. y Lyons, K. (2006). **BMP Signaling in the Cartilage Growth Plate**. Current topics in Developmental Biology. 76: 1-48.

Pupko, T., Pe'er, I., Shamir, R. y Graur, D. (2000). **A fast algorithm for joint reconstruction of ancestral amino acid sequences**. Molecular Biology and Evolution, 17(6):890–896.

Purdy, M.A. (2008). **Phylogenetics and Medicinal Protein engineering**. Capítulo 3 en Medicinal Protein Engineering. Editado por Yury E. Khudiyakov. CRC Press 2008. Print ISBN: 978-0-8493-7368-8. eBook ISBN: 978-1-4200-0730-5

Reddi, A. H.; Huggins, C. (1972). **Biochemical Sequences in the Transformation of Normal Fibroblasts in Adolescent Rat**. PNAS 69 (6): 1601–1605.

Reddi, AH.(1981). **Cell biology and biochemistry of endochondral bone development**. Collagen and related research. 1: 209–226.

Reddi, AH. 1998; **Cartilage-derived morphogenetic proteins and cartilage morphogenesis**. Microscopy Research and Technique. 43: 131–136.

Reddi, A.H. (2005). **BMPs: from bone morphogenetic proteins to body morphogenetic proteins**. Cytokine growth factor reviews. 16: 3, 249-250

Sampath, T. K.; Reddi, A. H. (1981). **Dissociative Extraction and Reconstitution of Bone Matrix Components Involved in Local Bone Differentiation**. PNAS 78 (12): 7599–7603.

Sampath, T.K., Muthukumar, N., Reddi, A.H. (1987). **Isolation of Osteogenin, an Extracellular Matrix-Associated Bone-Inductive Protein, by Heparin Affinity Chromatography**. PNAS 84 (20): 7109–7113

Seeherman, H., y Wozney, J.M. (2005). **Delivery of bone morphogenetic proteins for orthopaedic tissue regeneration**. Cytokine growth factor reviews. 16:329-345.

Senn, N. (1889). **On the healing of aseptic bone cavities by implantation of antiseptic decalcified bone**. American Journal of Medical Sciences. 98: 219–243.

Singhatanadgit, W. y Olsen, S.I. (2006). **Bone morphogenetic protein receptors and bone morphogenetic protein signaling are controlled by tumor necrosis factor- α in human bone cells**. The International Journal of Biochemistry & Cell Biology. 38 (10): 1794–1807.

Shi, M., Zhu, J., Wang, R., Chen, X., Mi, L., Walz, T y Springer, T.A. (2011). **Latent TGF- β structure and activation**. *Nature*. 474:343-351.

Spinella-Jaegle S, Roman-Roman S, Faucheu C et al. (2001). **Opposite effects of bone morphogenetic protein-2 and transforming growth factor-beta1 on osteoblast differentiation**. *Bone*, 29: 323-330.

Stackhouse, Presnell, McGeehan, Nambiar, & Benner. (1990). **The ribonuclease from an extinct bovid ruminant**. *FEBS Lett*. 262: 104–106.

Stewart, C.-B., J. W. Schilling, and A. C. Wilson. 1987. **Adaptive evolution in the lysozymes of foregut fermenters**. *Nature* 330:401-404.

Ström et al. 2011. **Burden of Fractures in France, Germany, Italy, Spain, Sweden, and the UK. Osteoporosis International**. Volume 22, Supplement 1 (2011), 91-95, DOI: 10.1007/s00198-011-1554-9 .

Swanson, K.W., Irwin, D.M. y Wilson, A.C.. 1991. **Stomach lysozyme gene of the Langur monkey tests for convergence and positive selection**. *Journal of Molecular Evolution*. 33:418-425.

Swofford, D.L. y Maddison, W.P. 1987. **Reconstructing Ancestral Character States under Wagner Parsimony**. *Mathematical Biosciences*, 87:199–229.

Tabas JA, Zasloff M, Wasmuth JJ et al. (1991). **Bone morphogenetic protein: chromosomal localization of human genes for BMP1, BMP2A, and BMP3**. *Genomics*, 9: 283-289.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, and Kumar S (2011) **MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods**. *Molecular Biology and Evolution* 28: 2731-2739.

Tamura K, Dudley J, Nei M & Kumar S. 2007. **MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0**. *Molecular Biology and Evolution* 24: 1596-1599.

Thornton, J. 2004. **Resurrecting ancient genes - experimental analysis of extinct molecules**. *Nature Reviews - Genetics* 5: 366-376.

Tobin, M.B., Gustafsson, C. y Huisman, G.W. 2000. **Directed evolution: the 'rational' basis for 'irrational' design**. *Current opinion in structural biology*. 10(4): 421-427.

Tou, L., Quibria, N. y Alexander, J.M. 2003. **Transcriptional regulation of the human Runx2/Cbfa1 gene promoter by bone morphogenetic protein-7**. *Molecular and Cellular Endocrinology*. 205: 121-129

Ugalde, J.A., B. Chang, S.W. y Matz, M.V. 2004. **Evolution of coral pigments recreated**. *Science*, 305(5689):1433–1433.

Urist, Marshall R. (1965). **Bone: formation by autoinduction**. *Science* 12:150 (698): 893–899.

Urist, Marshall R.; Strates, Basil S. (1971). **Bone Morphogenetic Protein**. Journal of Dental Research 1971 50 (6): 1392–1406.

Van den Heuvel, M.G.L. y Dekker, C. 2007. **Motor proteins at work for biotechnology**. Science, 317:333-336.

Wagner DO, Sieber C, Bhushan R, Börgermann JH, Graf D, y Knaus P., 2010. **BMPs: from bone to body morphogenetic proteins**. Science Signaling. 3:107.

Wang EA, Rosen V, Cordes P, et al. 1988. **Purification and characterization of other distinct bone-inducing factors**. Proc Natl Acad Sci USA. 85: 9484–9488.

Wikesjö et al., UME; Qahash, M; Huang, Y-H; Xiropaidis, A; Polimeni, G; Susin, C (2009). **Bone morphogenetic proteins for periodontal and alveolar indications; biological observations – clinical implications**. Orthodontics & Craniofacial Research 12 (3): 263–270.

White, A.P., Vaccaro, A.R., Hall, J.A., Whang, P.G., Friel, B.C., Mckee, M.D. 2007. **Clinical applications of BMP-7/OP-1 in fractures, nonunions and spinal fusion**. International Orthopaedics. 31:735-741.

Whitman, M. (August 1998). **Smads and early developmental signaling by the TGFbeta superfamily**. Genes Dev. 12 (16): 2445–62.

Wozney J.M., Rosen V., Celeste A.J. 1988. **Novel regulators of bone formation: molecular clones and activities**. Science, 242: 1528-1534.

Wozney, J.M. 2002. **Overview of Bone Morphogenetic Proteins**. SPINE Volume 27 (16S): S2–S8

World Health Organization. 2003. **The Burden of Musculoskeletal Conditions at the Start of the New Millennium**, Technical Report Series Number 919.

Yang, Z., Kumar, S. y Nei, M. 1995. **A new method of inference of ancestral nucleotide and amino acid sequences**. Genetics, 141(4):1641–1650.

Yang Z. 1997. **PAML: a program package for phylogenetic analysis by maximum likelihood**. Bioinformatics.13(5):555-556.

Yang Z. 2007. **PAML 4: phylogenetic analysis by maximum likelihood**. Molecular Biology and Evolution.24(8):1586-1591.

Zhang, J. y Nei, M, 1997. **Accuracies of ancestral amino acid sequences inferred by the Parsimony, Likelihood and Distance Methods**. *Journal of Molecular Evolution*, 44:S139-S146

Zimmermann G, Müller U, Löffler C, Wentzensen A, Moghaddam A (November 2007). **Therapeutic outcome in tibial pseudarthrosis: bone morphogenetic protein 7 (BMP-7) versus autologous bone grafting for tibial fractures**. Unfallchirurg 110 (11): 931–8.

7. CONCLUSIONES, RECOMENDACIONES Y PRODUCTOS DEL PRESENTE TRABAJO

CONCLUSIONES GENERALES

En el presente trabajo se analizaron aminoácidos individuales y en tripletas, junto con su relación a las estructuras secundarias que adopta.

La muestra utilizada en el análisis, PDBGlobal, es representativa de la totalidad de PDB y del universo proteico representado en él. Esta muestra, y las submuestras resultantes de la agrupación por criterios como contenido de estructura secundaria, localización celular, organismo del que provienen, Phylum al que pertenecen los organismos, Tipo de Enzima según la Enzyme Commission y Familia proteica, se ponen a disposición usarán en trabajos posteriores a este proyecto de investigación.

Se estudiaron las frecuencias de aminoácidos y de tripletas de aminoácidos tanto en la muestra PDBGlobal como en subconjuntos que agrupan las proteínas de acuerdo con diferentes criterios. Como resultados principales se obtuvieron:

- Las Cisteínas presentan los valores más bajos en la muestra de la totalidad de las proteínas, pero el contenido en estos residuos de aminoácidos, varía ampliamente entre los grupos de proteínas, indicando variados contenidos de puentes de disulfuro como opción para estabilizar la estructura tridimensional y la unión entre subunidades.
- La Cisteína aparece en frecuencias notoriamente elevadas en proteínas de organismos eucariotes, mientras que se presenta en cantidades notoriamente menores, en organismos procariotes, lo que indica que los primeros estabilizan sus proteínas soportándose más en puentes de disulfuro, y que los segundos estabilizan la estructura 3D de las proteínas, basándose en interacciones débiles.
- Las tripletas aparecen en frecuencias diferentes de lo esperado, calculando éstas últimas con base en las frecuencias de los aminoácidos que las componen, lo que implica que para estimar la probabilidad de ocurrencia de una secuencia proteica, es más preciso usar las frecuencias de tripletas de residuos, y que otras reglas rigen el uso de combinaciones de residuos en la construcción de las proteínas.
- Las frecuencias de aminoácidos y de tripletas en las proteínas agrupadas por tipo de enzima, presentan poca variabilidad, lo que sugiere que la clasificación estructural de enzimas, no se puede realizar con base en la actividad catalítica y tipos de reacciones que realizan.
- Las frecuencias de aminoácidos y tripletas son marcadamente diferentes entre familias de proteínas, y diferentes comparadas con la muestra del universo proteico, lo que indica que en

la clasificación estructural de proteínas es más adecuado si se consideran las proteínas agrupadas por familia proteica.

- La agrupación en un alfabeto reducido, evidencia que los aminoácidos hidrofóbicos son los más abundantes en las proteínas, posiblemente por su papel en la iniciación del plegamiento mediante el colapso hidrofóbico, y su lugar en el corazón hidrofóbico de proteínas globulares.
- Las proteínas de pequeño tamaño, tienden a estabilizar su estructura mediante interacciones débiles, dado que muestran contenidos de Cisteínas, especialmente bajos.

Se generalizaron los conceptos de N-cap y C-cap, al analizar las transiciones entre diferentes estructuras secundarias.

Al analizar aminoácidos individuales y tripletas de aminoácidos, con respecto a las frecuencias, con las que adoptan diferentes estructuras secundarias, se obtuvieron las siguientes conclusiones principales:

- El contenido de estructura secundaria varía entre grupos de proteínas, siendo las estructuras helicoidales, las estructuras más comunes, y la suma de los tres tipos de giros, las siguientes estructuras más comunes.
- Para construir un giro al azar, se recomienda usar combinaciones de los residuos Histidina y Prolina, dada su alta frecuencia, propensión y valor de IPE en estas estructuras secundarias.
- Las proteínas agrupadas por familia proteica y por contenido de estructura secundaria, presentan las mayores variaciones en estructura secundaria.
- Hay tripletas que aparecen casi exclusivamente en una estructura secundaria en particular, lo que indica que para hacer predicción de estructura, no se pueden hacer con aminoácidos individuales, y además, hay que mirar el contexto de secuencia.
- Las propensiones de las tripletas, muestran una parte del panorama que los aminoácidos individuales no permiten ver. Que algunas combinaciones de residuos (en este caso, tripletas), presentan propensiones altas hacia la formación de transiciones entre estructuras secundarias particulares.
- El estudio de tripletas permite identificar secuencias de aminoácidos frecuentes en las tripletas transicionales entre dos estructuras secundarias. Este análisis puede ser considerado como una generalización los conceptos de N-cap y C-cap, que flanquean las hélices.

Para profundizar el análisis de la relación entre aminoácidos y estructura secundaria, se emplearon tres indicadores: la propensión de Chou y Fasman, la entropía conformacional y el indicador de preferencia estructural, desarrollado por el autor de este proyecto. En cada caso se introdujo el concepto para aminoácidos individuales y se generalizaron las definiciones para hacerlas válidas para secuencias cortas de aminoácidos. De acuerdo con estos indicadores se puede decir:

- Los valores de IPE permiten identificar proteínas con una mayor tendencia a mantener la estructura secundaria (y por tanto, la terciaria), y también proteínas con baja tendencia a conservar la estructura secundaria. En el caso de proteínas estables como GFP, el IPE permite identificar que está formada por combinaciones más estables en su estructura secundaria, mientras que los priones están formados por combinaciones de residuos menos estables, detectables por sus valores de IPE.
- El alcance de la inferencia que se puede obtener de las propensiones, para la construcción de nuevas proteínas, resulta incompleto, pues no permite identificar las estructuras secundarias con tendencia a mantenerse estables. El valor de IPE, permite complementar el concepto de propensión, al identificar combinaciones de residuos que forman una estructura secundaria con tendencia a cambiar ó a conservarse.
- El IPE resulta casi opuesto al concepto de Entropía conformacional, en el caso de aminoácidos individuales. Sin embargo, al analizar secuencias cortas, las diferencias entre el IPE y las Entropías, se hace más marcado y ya no pueden usarse indistintamente para interpretar una secuencia proteica con su estructura secundaria. El IPE permite identificar proteínas ó zonas de proteínas, que pueden cambiar el plegamiento, según se puede inferir de la observación de proteínas como Priones.

En el marco de los estudios doctorales se hizo una pasantía entre los laboratorios de Química Física de Proteínas e Ingeniería de Proteínas de la Universidad de Granada, y el laboratorio de Biología Sintética de GeorgiaTech, durante la cual se hizo la reconstrucción ancestral de Proteínas Morfogenéticas de Hueso y enzimas Beta-Lactamasas. Resultados de esta pasantía son:

- Para construir proteínas con termoestabilidad magnificada y promiscuidad en su función, sería conveniente usar las metodologías de la reconstrucción ancestral.
- La termoestabilidad magnificada de las proteínas ancestrales correspondientes a períodos precámbricos, es coherente con la evidencia geológica que dice que el planeta estaba a mayores temperaturas que hoy en día.
- Las proteínas ancestrales presentan mayor promiscuidad que las proteínas de hoy, lo que es coherente con la teoría evolutiva.
- Las herramientas de la Reconstrucción de Secuencias Ancestrales permiten asomarse al pasado, para hacer inferencias evolutivas, sobre genes y proteínas, y eventualmente usar esa información para el rediseño de proteínas de hoy en día.

7.2 RECOMENDACIONES

Dado que los seis tipos de enzima de la comisión Enzimática, no mostraron variabilidad en las frecuencias ni propensiones, y que abarcan una gama amplia de reacciones químicas, recomendamos que se realice este trabajo con las subclases de enzimas, que realizan reacciones químicas más específicas.

Extender el estudio de tripletas a la sustituibilidad de las mismas, para la construcción de matrices de sustitución de aminoácidos, para la globalidad del PDB, y por grupos de proteínas.

Explorar el uso del IPE en genomas ó proteomas completos y verificar experimentalmente su poder predictor de proteínas con tendencias a modificar su plegamiento.

Dado que las familias proteicas son las que presentan mayor variabilidad en las frecuencias de aminoácidos y tripletas, y en los valores de propensión, entropía e IPE, pueden considerarse un interesante reservorio de información para hacer inferencia sobre las combinaciones de aminoácidos y las condiciones particulares de las proteínas, que determinan la estructura funcional de una proteína. Por ello, se recomienda extender este trabajo a un número mayor de familias proteicas, de manera que se obtenga una mayor variedad de características particulares que permitan inferir sobre su estructura y función.

Recomendamos extender este trabajo con tripletas, teniendo en consideración los ángulos Phi y Psi como medio de caracterizar las tripletas y las estructuras en que aparecen.

Sería interesante extender el presente trabajo a la estructura tridimensional comparando tripletas en diferentes proteínas, mediante el uso de los valores de RMSD (Root Mean-Square Deviation), que permitan inferir sobre las tripletas tridimensionalmente más estables ó sobre las conformaciones que pueden adquirir.

Se recomienda extender la metodología aplicada en el presente trabajo, a la caracterización de estructuras secundarias con características particulares, como hélices anfipáticas, voluminosas, completamente hidrofóbicas, etc.

7.3. PRODUCTOS Y PUBLICACIONES

7.3.1. Del trabajo en proteínas del PDB

Publicaciones:

Se publicaron dos capítulos de libro:

1. **Mejía-Carmona, D.F., Tischer, I. Design of a new indicator of structural preference for residue triplets into proteins.** Proceedings, International Seminar on Bioinformatics and Systems Biology, Popayán, Colombia. **ISBN 978-958-9451-21-2** Fecha de publicación: Marzo, 2007.
2. Mossos, N., Mejia-Carmona, D.F., Tischer, I. 2014. **FS-Tree: Sequential Association Rules and First Applications to Protein Secondary Structure Analysis.** Advances in Computational Biology - Advances in Intelligent Systems and Computing. 232: 189-198. **ISBN 978-3-319-01568-2.** Capítulo que se puede consultar en http://link.springer.com/chapter/10.1007%2F978-3-319-01568-2_28 y libro, consultable en: http://download.springer.com/static/pdf/653/bfm%253A978-3-319-01568-2%252F1.pdf?auth66=1383966956_162bfca7def6d969e79e88e268cf3ba1&ext=.pdf

Presentaciones en eventos locales, nacionales e internacionales:

1. Se presentó una ponencia oral en el simposio internacional “**2nd International Seminar on Genomics, Proteomics, Bioinformatics and Systems Biology -2006**”.
2. Se presentó una ponencia oral en el congreso nacional “**II Congreso Colombiano de Biología Computacional y Bioinformática, 2013**”.: Mossos, N., Mejia-Carmona, D.F., Tischer, I. **FS-Tree: Sequential Association Rules and First Applications to Protein Secondary Structure Analysis.** http://www.ualdas.edu.co/index.php?option=com_content&view=article&catid=419%3Auniversidad-al-dia&id=8492%3Aii-congreso-colombiano-de-biologia-computacional-y-bioinformatica&Itemid=1039.
3. Se presentaron ponencias orales en tres eventos locales de divulgación de resultados de trabajos científicos: **Semana de la ingeniería, Universidad del Valle, versiones 2006 y 2008**), y **Simposio Grupo de Estudio y Trabajo en Genética, GETEG, 2007**.

7.3.2. Del rediseño de proteínas por reconstrucción ancestral

Publicaciones:

Se publicó un artículo científico en una revista internacional de prestigio (la revista de mayor impacto en Química): Journal of the American Chemical Society.

1. Risso, V., Gaviria, JA., **Mejia-Carmona, DF.**, Gaucher, E., Sanchez-Ruiz, JM. **Hyperstability and enhanced promiscuity in resurrected Precambrian β -Lactamases.** Journal of the American Chemical Society. Feb, 2013.

Este artículo fue comentado en numerosas páginas de ciencia en el mundo, como la American Chemical Society, Sciencedaily, Sciencenews, y otros, y alcanzó el quinto lugar entre los artículos de mayor interés en la entidad "Faculty of 1000":

<http://wildtypes.wordpress.com/2013/03/04/ancient-enzymes-conferring-antibiotic-resistance-re-created-in-the-lab/>

<http://www.sciencedaily.com/releases/2013/02/130227102028.htm>

http://portal.acs.org/portal/acs/corg/content?nfpb=true&pageLabel=PP_ARTICLEMAIN&node_id=223&content_id=CNBP_032276&use_sec=true&sec_url_var=region1&uuid=09fe5a45-35af-4ccf-9e0e-bd886db74c7e

http://www.science20.com/news_articles/3_billionyearold_antibioticresistant_proteins_resurrected-104944

http://zeenews.india.com/news/science/three-billion-year-old-proteins-resurrected_832121.html

http://articles.timesofindia.indiatimes.com/2013-02-28/health/37351145_1_antibiotic-resistance-proteins-beta-lactamase

<http://www.sciencenewslines.com/summary/2013022717350003.html>

<http://www.deccanherald.com/content/315547/three-billion-year-old-proteins.html>

<http://m.phys.org/news/2013-02-resurrection-billion-year-old-antibiotic-resistance-proteins.html>

<http://f1000.com/prime/717979625>

<http://www.sci-news.com/biology/article00904.html>

http://zeenews.india.com/news/science/three-billion-year-old-proteins-resurrected_832121.html

http://www.science20.com/news_articles/3_billionyearold_antibioticresistant_proteins_resurrected-104944

http://articles.timesofindia.indiatimes.com/2013-02-28/health/37351145_1_antibiotic-resistance-proteins-beta-lactamase

<http://www.sciencenewslines.com/summary/2013022717350003.html>

<http://www.deccanherald.com/content/315547/three-billion-year-old-proteins.html>

<http://m.phys.org/news/2013-02-resurrection-billion-year-old-antibiotic-resistance-proteins.html>

<http://f1000.com/prime/717979625>

<http://www.sci-news.com/biology/article00904.html>

<http://wildtypes.wordpress.com/2013/03/04/ancient-enzymes-conferring-antibiotic-resistance-re-created-in-the-lab/>

<http://pubs.acs.org/stoken/presspac/presspac/abs/10.1021/ja311630a>

<http://www.sciencedaily.com/releases/2013/02/130227102028.htm>

http://portal.acs.org/portal/acs/corg/content?nfpb=true&pageLabel=PP_ARTICLEMAIN&node_id=223&content_id=CNBP_032276&use_sec=true&sec_url_var=region1&uuid=09fe5a45-35af-4ccf-9e0e-bd886db74c7e

Presentaciones en eventos locales, nacionales e internacionales:

2. Se presentó un póster en el congreso internacional de la FEBS (Federation of European Biochemical Societies), en San Petersburgo, Rusia:
Risso, V., Gaviria, JA., **Mejía-Carmona, DF.**, Gaucher, E., Sanchez-Ruiz, JM. **Precambrian Antibiotic Resistance**. Federation of European Biochemical Societies – Congress 2013 July, 2013.
3. Se presentó una ponencia oral en un evento de Ingeniería de proteínas en España:
Mejía-Carmona, DF. **"Reconstructing ancestral growth factors"** - First Meeting on Protein Design Technologies - September 16th-18th (2011) - Castillo de Magalia, Navas del Marqués. Ávila. España. <http://prodestech.cib.csic.es/wp-content/uploads/2011/09/ProdestechProgramFINAL.pdf>
4. Se presentó una ponencia oral en el congreso internacional: International congress on Basic Biomedical Sciences. Universidad Libre, Cali, Colombia, Ago 2013. **Uso de la evolución molecular para rediseñar Beta-Lactamasas para la investigación en resistencia a antibióticos.**
5. Se presentó una ponencia oral en un evento local, semana de la investigación en química, Universidad Industrial de Santander, UIS, Bucaramanga Dic 6,2013.
6. Se presentó una ponencia oral en un evento local: **Rediseño de proteínas mediante la reconstrucción ancestral, aplicado a BMPs**. Semana de la química, Universidad Santiago de Cali, Octubre 31 de 2013.

Otros - Aplicaciones a patentes:

Se prepara una aplicación a una patente por las Beta-Lactamasas con promiscuidad y termoestabilidad elevada, para su uso en la evaluación de múltiples moléculas candidatas a ser usadas como antibióticos, por parte de la industria farmacéutica. Esta aplicación se presentará a la Oficina Española de Marcas y Patentes (OEMP).

Se están realizando los ensayos in vitro, e in vivo, en el grupo de Ingeniería de Tejidos de la Universidad Complutense de Madrid (<http://portal.ucm.es/web/ingenieria-de-tejidos/inicio>), para aplicar a una patente con las Proteínas Morfogenéticas de Hueso, y su uso en cirugías ortopédicas y maxilofaciales. Esta aplicación se presentará inicialmente a la Oficina Española de Marcas y Patentes (OEMP).

Reporte de Faculty of 1000, mostrando el escalafón de artículos de mayor interés en Biología y Química en Febrero de 2013, con el artículo sobre Resurrección de Lactamas ancestrales, publicado en JACS, en el lugar 5 del “top-10” de artículos publicados en este campo.

Chemical Biology Article Rankings: Current Top 10 – F1000

//f1000.com/prime/rankings/topten/current/faculty/chembiol

GT-TimeOut GT Bio Portal GT Online Training Course Surveys GT Transcripts TechWorks Web Stats GT Library TechWorks CoS Finance CayuseGrants New JME Old JME

Article Recommendations Rankings F1000Prime Reports F1000 Faculty Blog MyF1000 Sign out Subscribe

ARTICLE RANKINGS
All Time Top 10
Current Top 10
All Time Most Viewed
Current Most Viewed
Hidden Jewels

BROWSE BY FACULTY
All | Biology | Medicine

Anesthesiology & Pain Management
Biochemistry
Bioinformatics & Computational Biology
Biotechnology
Cancer Biology
Cardiovascular Biology
Cardiovascular Disorders
Cell Biology
Chemical Biology
Critical Care & Emergency Medicine
Dermatology
Developmental Biology
Diabetes & Endocrinology
Ecology
Evolutionary Biology
Gastroenterology & Hepatology
Gastrointestinal Biology
Genomics & Genetics
Hematology
Immunology
Infectious Diseases
Metabolic & Endocrine Science
Microbiology
Molecular Biology
Molecular Medicine
Nephrology
Neurological Disorders
Neuroscience
Oncology
Ophthalmology
Otolaryngology
Pharmacology & Drug Discovery
Physiology
Plant Biology
Psychiatry
Public Health & Epidemiology
Renal Biology
Research Methodology

F1000Prime > Rankings > Articles: Current Top 10 >

Chemical Biology

Current Top 10 rankings are generated every day using only the articles recommended in F1000Prime during the preceding 14 days. Articles are ranked according to their total star rating, which is calculated from all the stars awarded to that article by the Faculty. If any articles have the same total rating, then the most recently added article is listed first.

☐ Save to MyF1000/Follow

☐ **12** **Principles for designing ideal protein structures.**
Koga N, Talsum-Koga R, Liu G, Xiao R, Acton TB, Monteleone GT, Baker D. Nature. 2012 Nov 8; 491(7423):222-7
[Abstract](#) | [Related Article Recommendations](#)
6 Recommendations
Charlotte Deane and Jamie Hill 27 Feb 2013 | Zaida R Wasserman | Reinhard Stelmer and Rainer Merk | Colin Kleanthous and Nick Housden | Xiyang Qiu and Ye Che | Nikolay Dokholyan and David Shiryanyants

☐ **9** **Genomic responses in mouse models poorly mimic human inflammatory diseases.**
Seok J, Warren HS, Cuenca AG, Mindrinos MN, ... Davis RW, Xiao W, Tompkins RG. Inflammation and Host Response to Injury, Large Scale Collaborative Research Program. Proc Natl Acad Sci U S A. 2013 Feb 26; 110(8):3507-12
[Abstract](#) | [Related Article Recommendations](#)
4 Recommendations
Linda Ramanan 05 Mar 2013 | Luc Teyton | John Lowe | Mike Klimkowski

☐ **3** **Inside the mind of a medicinal chemist: The role of human bias in compound prioritization during drug discovery.**
Kulchukhan PS, Vasilyeva NY, Xu J, Lindvall MK, Dillon MP, Glick M, Coley JD, Broojmans N. PLoS ONE. 2012; 7(11):e48476
[Abstract](#) | [Related Article Recommendations](#)
2 Recommendations
Jürgen Bajorath 28 Feb 2013 | Martin Stahl

☐ **3** **Late-Stage Deoxyfluorination of Alcohols with PhenoFluor.**
Sladojevich F, Arlow SI, Tang P, Ritter T. J Am Chem Soc. 2013 Feb 20; 135(7):2470-3
[Abstract](#) | [Related Article Recommendations](#)
1 Recommendation
Justin Chalker 26 Feb 2013

☐ **2** **Hyperstability and substrate promiscuity in laboratory resurrections of Precambrian β -lactamases.**
Risso VA, Gavira JA, Mejia-Carmona DF, Gaucher EA, Sanchez-Ruiz JM. J Am Chem Soc. 2013 Feb 27; 135(8):2899-902
[Abstract](#) | [Related Article Recommendations](#)
1 Recommendation
Justin Chalker 01 Mar 2013

F1000Research
The fastest way to publish

COMING SOON: NEXT-GEN PROTEIN PURIFICATION THAT SCALES TO FIT


BIO-RAD
ADVERTISEMENT

The F1000.com website uses cookies. By continuing to browse the site, you are agreeing to our use of cookies. [Find out more >](#)

Hyperstability and Substrate Promiscuity in Laboratory Resurrections of Precambrian β -Lactamases

Valeria A. Risso,^{*,†} Jose A. Gavira,[‡] Diego F. Mejia-Carmona,^{†,||} Eric A. Gaucher,[§] and Jose M. Sanchez-Ruiz^{*,†}

[†]Facultad de Ciencias, Departamento de Química Física, Universidad de Granada, 18071 Granada, Spain

[‡]Laboratorio de Estudios Crystallográficos, Instituto Andaluz de Ciencias de la Tierra (Consejo Superior de Investigaciones Científicas—Universidad de Granada), Avenida de las Palmeras 4, 18100 Armilla, Granada, Spain

[§]School of Biology, School of Chemistry, and Parker H. Petit Institute for Bioengineering and Biosciences, Georgia Institute of Technology, Atlanta, Georgia 30332, United States

S Supporting Information

ABSTRACT: We report a sequence reconstruction analysis targeting several Precambrian nodes in the evolution of class-A β -lactamases and the preparation and experimental characterization of their encoded proteins. Despite extensive sequence differences with the modern enzymes (~ 100 amino acid differences), the proteins resurrected in the laboratory properly fold into the canonical lactamase structure. The encoded proteins from 2–3 billion years (Gyr)-old β -lactamase sequences undergo cooperative two-state thermal denaturation and display very large denaturation temperature enhancements (~ 35 °C) relative to modern β -lactamases. They degrade different antibiotics in vitro with catalytic efficiencies comparable to that of an average modern enzyme. This enhanced substrate promiscuity is not accompanied by significant changes in the active-site region as seen in static X-ray structures, suggesting a plausible role for dynamics in the evolution of function in these proteins. Laboratory resurrections of 2–3 Gyr-old β -lactamases also endowed modern microorganisms with significant levels of resistance toward a variety of antibiotics, opening up the possibility of performing laboratory replays of the molecular tape of lactamase evolution. Overall, these results support the notions that Precambrian life was thermophilic and that proteins can evolve from substrate-promiscuous generalists into specialists during the course of natural evolution. They also highlight the biotechnological potential of laboratory resurrection of Precambrian proteins, as both high stability and enhanced promiscuity (likely contributors to high evolvability) are advantageous features in protein scaffolds for molecular design and laboratory evolution.

Antibiotic resistance genes are widespread in natural environments, partly as a result of human activities such as the use of antibiotics in agriculture and fish farming.¹ In all likelihood, however, antibiotic resistance existed before the modern use of antibiotics. In fact, resistance genes have been found in unpolluted sites such as remote Alaskan soil,^{2a} the bottom of the Pacific Ocean,^{2b} and even 30 000-year-old

permafrost sediments.^{2c} Furthermore, β -lactamases, the enzymes responsible for the primary mechanism of resistance toward β -lactam antibiotics, are widely distributed throughout the bacterial domain of life and are believed to have originated billions of years (Gyr) ago.³ The role antibiotics and antibiotic resistance play in modern natural environments is not fully understood,¹ so one can only speculate about the reasons for the existence of β -lactamases that are several billion years old. It has been suggested that they originated as a mechanism of defense to stave off competitors^{3a} or had a role in signaling.⁴

Laboratory resurrection of ancestral proteins⁵ has been shown to provide insight into the ancient properties of biomolecules as well as the intracellular and extracellular environments hosting these proteins. Here we report the derivation of statistically probable sequences for several Precambrian β -lactamases, the preparation in the laboratory of the corresponding proteins, and their experimental characterization in terms of stability, function, and structure.

We used a set of chromosomal sequences of extant class-A β -lactamases [Tables S1 and S2 and Figure S1 in the Supporting Information (SI)] to construct a phylogenetic tree encompassing Gram-positive and Gram-negative bacteria (Figures S2 and S3). The sequences were derived from a search in the NCBI database of complete genomes, and thus, plasmidic lactamases were not included. To avoid potential complications during the sequence reconstruction process that could arise from recent evolution during the antibiotic era, we determined that none of the sequences came from clinical isolates. The 75 sequences selected provide uniform coverage of the phyla in Bacteria, including the Proteobacteria, Actinobacteria, Firmicutes, and Cytophaga–Flexibacter–Bacteroides (CFB) lineages, and the topology of the inferred phylogenetic tree is sufficiently close to an accepted phylogeny of these organisms⁶ (Figure S4) to allow us to target well-defined Precambrian nodes (for details, see Materials and Methods in the SI). We focused our sequence reconstructions of β -lactamases on the last common ancestors (CAs) of enterobacteria (ENCA), Gammaproteobacteria (GPBCA), various Gram-negative bacteria (GNCA), and various Gram-positive and Gram-negative bacteria (PNCA). Estimates of divergence times⁶

Received: November 28, 2012

Published: February 9, 2013



indicate that these ancestors inhabited Earth about 1, 1.5, 2, and 3 Gyr ago, respectively. We used Bayesian statistics to reconstruct the sequences of the β -lactamases corresponding to these Precambrian nodes and accounted for the protein corresponding to the most probabilistic sequence (i.e., the sequence with the most probable amino acid at each position) for each phylogenetic node of interest (Figure S5). The reconstructed proteins have substantial amino acid sequence differences relative to extant β -lactamases and also to a consensus sequence computed from extant β -lactamases (Table S3). The phenotypic robustness of the laboratory resurrection was supported by measurements on additional sequences at the GNCA node derived from a Monte Carlo sampling of the posterior probability distribution (Figures S6–S10 and Table S4).

The laboratory resurrections of the ancestral β -lactamases were characterized to determine their structural, physicochemical, and functional properties (see Materials and Methods). Crystallization and 3D structure determination of the lactamases corresponding to the GNCA and ENCA nodes showed that they share the extant lactamase fold despite extensive sequence differences (122 and 55 amino acid differences, respectively, relative to TEM-1 β -lactamase⁷ (Figure 1). Closer inspection revealed that no important conformational changes reside in the

ancestral GNCA and ENCA structures relative to the TEM-1 lactamase structure [root-mean-square deviations (RMSDs) of 1.44 and 0.63 Å, respectively] (Figures 1 and S16). Minor movements having $\text{RMSD} \geq 2$ Å appear in the $\alpha\beta$ domain of the GNCA lactamase, corresponding to solvent-exposed loops and the terminal helix, but such movements can be expected. No substantial differences were found in the α domain, and all of the active-site amino acids occupy canonical space (RMSDs of 0.42 and 0.40 Å for GNCA and ENCA, respectively).

The thermal stability of the resurrected Precambrian β -lactamases was studied by differential scanning calorimetry (DSC). The resurrected 2–3 Gyr-old β -lactamases were found to be highly stable, as revealed by their denaturation temperatures (T_m), which were ~ 35 °C higher than those of extant lactamases, including that from the thermophilic *Bacillus licheniformis* (Figures 2A and S14). Also, the effect of scan rate

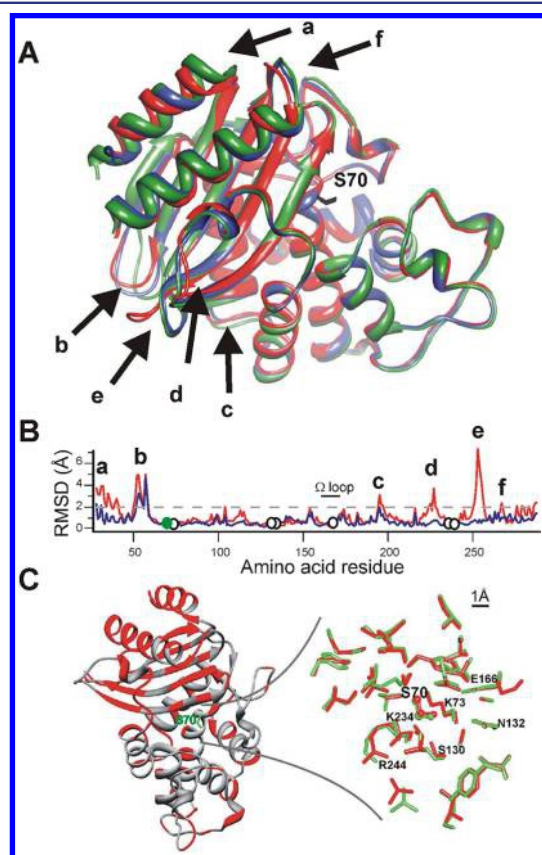


Figure 1. (A) Structural comparison of the TEM-1 β -lactamase (PDB entry 1btl; green) and the resurrected GNCA (PDB entry 4b88; red) and ENCA (PDB entry 3zdj; blue) β -lactamases. The catalytic serine S70 is shown. (B) RMSD plot. Circles represent the active-site positions (green for S70). Minor structural differences (labeled a–f) are seen only in the $\alpha\beta$ domain. (C) (left) GNCA lactamase structure. Positions with differences relative to modern TEM-1 β -lactamase are shown in red. (right) Side-chain conformations within a sphere of 8 Å centered at S70 for GNCA (red) and TEM-1 (green) β -lactamases. Active-site residues are labeled.

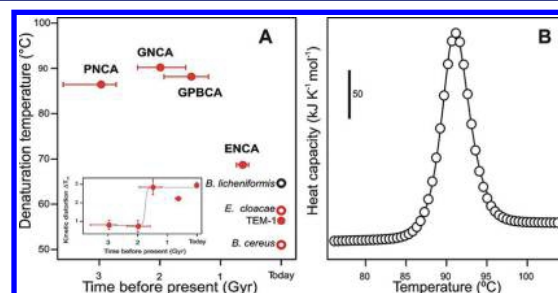


Figure 2. (A) Plot of denaturation temperature vs geological time for β -lactamases, with uncertainties in geological time from ref 6 (see Figure S14 for details). Inset: kinetic distortions in the thermal denaturation of β -lactamases as measured by the effect of scan rate on the DSC profiles (see Figure S15 for details). (B) DSC profile for thermal denaturation of the resurrected GNCA β -lactamase: O, experimental data; solid line, best fit using the two-state equilibrium model (see Figure S15 for details and additional fits).

on the DSC profiles indicated that kinetic distortions are substantially smaller for the resurrected ancestral lactamases (Figure 2A inset and Figure S15), suggesting enhanced kinetic stabilization against irreversible denaturation processes.⁸ In fact, the DSC profiles for the resurrections of 2–3 Gyr-old β -lactamases are well-described by a simple native-unfolded, two-state equilibrium model (Figures 2B and S15).

It is unlikely that the enhanced stability shown in Figure 2A was obtained by chance or through biases associated with the sequence reconstruction process because the sequence differences are extensive (~ 100 amino acid replacements) and, as is well-known, most mutations in a protein are destabilizing.^{9,10} Moreover, the T_m enhancement of ~ 35 °C is substantially larger than those reported in stability engineering studies of modern β -lactamases. For instance, combinatorial recruitment of consensus mutations led to a T_m enhancement of 9 °C for the β -lactamase of *Enterobacter cloacae*,¹¹ and consensus variants of TEM-1 β -lactamase obtained through intense neutral drifts displayed T_m enhancements of up to 7 °C.¹² In fact, the highest T_m reported to date for a variant of TEM-1 β -lactamase is 69 °C,¹³ which is still ~ 20 °C below the T_m values for the resurrected 2–3 Gyr-old β -lactamases (Figure 2A).¹⁴ It is also noteworthy that two other recent Precambrian protein resurrection studies^{5d,e} involving thioredoxins and elongation factors also found T_m enhancements of 30–35 °C upon “traveling back in time” 2–3 Gyr. The fact that thioredoxins, elongation factors, and β -lactamases differ in structure, size, and function strongly suggests that hyperstability

should be a general outcome for many Precambrian resurrection efforts, likely reflecting the adaptation of Precambrian proteins to the hot conditions of the ancestral oceans.

In vitro measurements of the catalytic properties for the hydrolysis of several antibiotics (Figure S11) revealed that the resurrections of 2–3 Gyr-old β -lactamases have somewhat decreased values of $k_{\text{cat}}/K_{\text{m}}$ for penicillin antibiotics but substantially enhanced values for third-generation antibiotics (Figures 3A and S12 and Table S5). Actually, while the extant

from clinical isolates has sometimes been linked to significant structural alterations.¹⁶ Thus, it is surprising that the structures determined for the promiscuous GNCA β -lactamase and the modern TEM-1 penicillin specialist did not show significant differences in the active-site region (Figure 1C) or the critical Ω loop (Figure 1B). While subtle structural effects could certainly play a role,¹⁷ it appears plausible that the evolution of function in this system is related to dynamic effects^{18,19} that may not be apparent in the static X-ray structures.^{20,21}

To assess the capability of each resurrected ancestral β -lactamase to confer antibiotic resistance within a modern organism, we transformed a DH5 α *Escherichia coli* strain with a PBR322 plasmid containing a reconstructed gene sequence encoding a resurrected protein. To minimize alterations in protein export to the periplasm, the reconstructed ancestral sequence was fused to the 23-residue peptide signal sequence from the extant *E. coli* lactamase. Minimum inhibitory concentrations (MICs) were subsequently determined for the penicillin-family antibiotic ampicillin and also for several third-generation antibiotics (Figure 3D and Table S6). The resurrected β -lactamases showed decreased MICs for ampicillin (compared with the extant β -lactamase from *E. coli*) but enhanced MICs for other antibiotics, in qualitative agreement with the in vitro catalytic efficiencies described above (compare panels A and D in Figure 3).²² The capability of the resurrected Precambrian β -lactamases to confer antibiotic resistance within a modern organism opens up the interesting possibility of performing directed laboratory evolution of these proteins in a straightforward manner (i.e., through in vivo selection for enhanced resistance).

Several published studies of laboratory resurrection of ancestral proteins have provided narratives describing evolutionary adaptations over planetary time scales. One plausible narrative suggested by the present results is that the enhanced stability of our resurrected β -lactamases reflects the adaptation to high temperatures predicted to be associated with Earth's early oceans that hosted life.^{5d,e} Consequently, the decrease in stability over several Gyr (Figure 2) may reflect the cooling of the oceans, although the existence of stability/activity trade-offs cannot be ruled out. The resurrected proteins' enzymatic promiscuity may be related to the wide variety of substances Precambrian lactamases had to hydrolyze^{3a} or may simply reflect generalist properties of the first stages in the evolution of lactamase enzymes.²³ This generalist property is revealed by the capability of our resurrections to hydrolyze different types of antibiotics, including modern third-generation ones (Figure 3A,D). Their functional conversion from generalists to penicillin specialists over a several-Gyr time scale (Figure 3) could be an adaptation to the development of efficient penicillin-like antibiotics by other organisms (interestingly, the divergence time of fungi is ~ 1.2 Gyr before present).⁶ The overall enzymatic and physicochemical characterization of the resurrected β -lactamases thus supports the notions that Precambrian life was thermophilic and that proteins can evolve from generalists (displaying substrate promiscuity) to specialists capable of efficient enzymatic turnover during the course of natural evolution.

The availability of resurrected Precambrian β -lactamases opens up new possibilities in the study of the emergence of antibiotic resistance. For instance, it should now be possible to perform laboratory replays of the molecular tape of lactamase evolution to probe the molecular determinants of the efficiency of lactamases in adapting to different types of antibiotics. From a different perspective, we point out that both high stability and

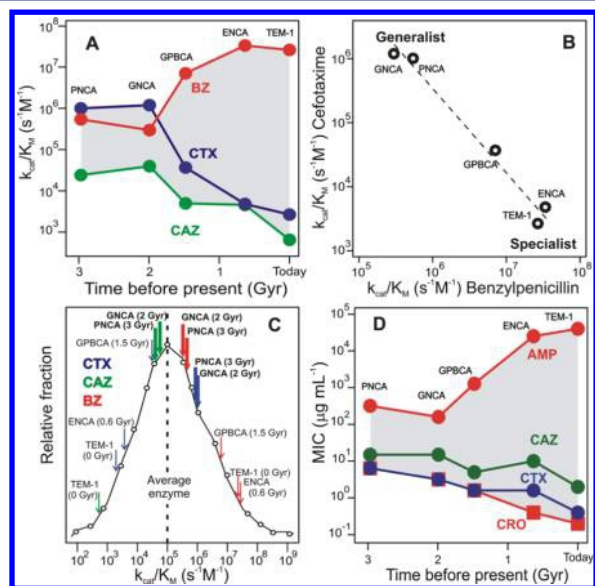


Figure 3. (A) $k_{\text{cat}}/K_{\text{m}}$ for the in vitro hydrolysis of antibiotics catalyzed by extant and resurrected β -lactamases. (B) Generalist-to-specialist conversion as illustrated by the catalytic efficiencies for benzylpenicillin (BZ) and cefotaxime (CTX). (C) Distribution of catalytic efficiencies. Open circles represent values for modern enzymes from ref 15, and arrows indicate values for the β -lactamases studied in this work. Thicker and longer arrows are used for the substrate-promiscuous laboratory resurrections of the 2–3 Gyr-old β -lactamases. (D) MICs determined using an *E. coli* strain transformed with plasmids containing the resurrected β -lactamases. In (A) and (D), shading is meant to illustrate the change in the range of values with geological time. The uncertainties in geological time are given in Figure 1A. The Michaelis–Menten parameters and MICs (together with the associated errors) are collected in Tables S5 and S6. BZ, ampicillin (AMP), and the third-generation antibiotics CTX, ceftazidime (CAZ), and ceftriaxone (CRO) (Figure S11) were studied.

TEM-1 β -lactamase is clearly a penicillin specialist, the GNCA and PNCA β -lactamases show roughly similar levels of catalytic efficiency for the different antibiotics studied (penicillin and third-generation) and can be regarded as substrate-promiscuous enzymes (Figure 3B). Bar-Even et al.¹⁵ recently analyzed the catalytic efficiencies for several thousand (extant) enzymes and found that the “average enzyme” has a moderate efficiency of $k_{\text{cat}}/K_{\text{m}} \sim 10^5 \text{ M}^{-1} \text{ s}^{-1}$, which is much lower than the diffusion limit of 10^8 – $10^9 \text{ M}^{-1} \text{ s}^{-1}$. Figure 3C shows the distribution of catalytic efficiency for extant enzymes reported in ref 15 with some added arrows that locate the catalytic efficiency levels of the β -lactamases studied here (see Figure S13 for the K_{m} and k_{cat} distributions). The efficiencies of the GNCA and PNCA β -lactamases toward the different antibiotics used are similar to that of the average extant enzyme. The resurrected 2–3 Gyr-old β -lactamases can thus be described as moderately efficient promiscuous enzymes. Broad substrate scope in lactamases

promiscuity (likely contributors to high evolvability) are advantageous features in protein scaffolds for molecular design and laboratory evolution. High stability allows destabilizing but functionally useful mutations to be accepted without compromising proper folding.^{24a} Substantial promiscuity is essential for laboratory evolution of high levels of a targeted function.^{24b} The reported simultaneous occurrence of both features thus supports the potential of laboratory resurrection of Precambrian proteins in protein engineering and protein biotechnology.

■ ASSOCIATED CONTENT

■ Supporting Information

Experimental details and additional experimental data. This material is available free of charge via the Internet at <http://pubs.acs.org>. Atomic coordinates and structure factors for the reported crystal structures have been deposited with the Protein Data Bank as entries 4b88 and 3zjdj.

■ AUTHOR INFORMATION

Corresponding Author

vrisso@ugr.es; sanchezr@ugr.es

Present Address

[†]Postgrado en Ciencias Biomédicas, Facultad de Salud, Universidad del Valle, sede San Fernando, Cali, Colombia 760042.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by Grants BIO2009-09562, CSD2009-00088 (J.M.S.-R.), BIO2010-16800, and “Factoría Española de Crystalización”, Consolider-Ingenio 2010 (J.A.G.) from MICINN, Spain, FEDER Funds (J.M.S.-R. and J.A.G.) and grants from NASA Exobiology (NNX08AO12G), the NASA Astrobiology Institute, and Georgia Tech (E.A.G.). The authors thank Dr. Beatriz Ibarra-Molero, Dr. Asunción Delgado-Delgado, Dr. Rocío Arco-Gonzalez, and Dr. Zi-Ming Zhao for advice and suggestions and Dr. Inés G. Muñoz for help with X-ray data collection at CNIO.

■ REFERENCES

- (1) Allen, H. K.; Donato, J.; Wang, H. H.; Cloud-Hansen, K. A.; Davies, J.; Handelsman, J. *Nat. Rev. Microbiol.* **2010**, *8*, 251.
- (2) (a) Allen, H. K.; Moe, L. A.; Rodbumer, J.; Gaarder, A.; Handelsman, J. *ISME J.* **2009**, *3*, 243. (b) Toth, M.; Smith, C.; Frase, H.; Mobashery, S.; Vakulenko, S. *J. Am. Chem. Soc.* **2010**, *132*, 816. (c) D’Costa, V. M.; King, C. E.; Kalan, L.; Morar, M.; Sung, W. W. L.; Schwartz, C.; Froese, D.; Zazula, G.; Calmels, F.; Debruyne, R.; Golding, G. B.; Poinar, H. N.; Wright, G. D. *Nature* **2011**, *477*, 457.
- (3) (a) Hall, B. G.; Barlow, M. *Drug Resist. Updates* **2004**, *7*, 111. (b) Wright, D. G. *Nat. Rev. Microbiol.* **2007**, *5*, 175. (c) Kim, K. M.; Caetano-Anollés, G. *BMC Evol. Biol.* **2011**, *11*, 140.
- (4) Yim, G.; Wang, H. H.; Davies, J. *Philos. Trans. R. Soc., B* **2007**, *362*, 1195.
- (5) (a) Benner, S. A.; Sassi, S. O.; Gaucher, E. A. *Adv. Enzymol. Relat. Areas Mol. Biol.* **2007**, *75*, 1. (b) Thornton, J. W. *Nat. Rev. Genet.* **2004**, *5*, 366. (c) Carroll, S. M.; Ortlund, E. A.; Thornton, J. W. *PLoS Genet.* **2011**, *7*, No. e1002117. (d) Gaucher, E. A.; Govindarajan, S.; Ganesh, O. K. *Nature* **2008**, *451*, 704. (e) Perez-Jimenez, R.; Inglés-Prieto, A.; Zhao, Z.-M.; Sanchez-Romero, I.; Alegre-Cebollada, J.; Kosuri, P.; Garcia-Manyes, S.; Kappock, T. J.; Tanokura, M.; Holmgren, A.; Sanchez-Ruiz, J. M.; Gaucher, E. A.; Fernandez, J. M. *Nat. Struct. Mol. Biol.* **2011**, *18*, 592.
- (6) *The Timetree of Life*; Hedges, S. B., Kumar, S., Eds.; Oxford University Press: New York, 2009.
- (7) Jelsch, C.; Mourey, L.; Masson, J. M.; Samara, J. P. *Proteins* **1993**, *16*, 364.
- (8) Sanchez-Ruiz, J. M. *Biophys. Chem.* **2010**, *148*, 1.
- (9) Godoy-Ruiz, R.; Ariza, F.; Rodriguez-Larrea, D.; Perez-Jimenez, R.; Ibarra-Molero, B.; Sanchez-Ruiz, J. M. *J. Mol. Biol.* **2006**, *362*, 966.
- (10) Tokuriki, N.; Stricher, F.; Schymkowitz, J.; Serrano, L.; Tawfik, D. S. *J. Mol. Biol.* **2007**, *369*, 1318.
- (11) Main, N.; Liu, A. D.; Ramer, S.; Ahle, W.; Meijer, D.; Metin, M.; Wong, S.; Gualfetti, P.; Schellenberger, V. *Protein Eng.* **2004**, *17*, 787.
- (12) Bershtein, S.; Goldin, K.; Tawfik, D. S. *J. Mol. Biol.* **2008**, *379*, 1029.
- (13) Kather, I.; Jakob, R. P.; Dobbek, H.; Schmid, F. X. *J. Mol. Biol.* **2008**, *383*, 238.
- (14) Also, a T_m enhancement of $\sim 35^\circ\text{C}$ is larger than those typically achieved by rational design of protein stability. For instance, a 10°C enhancement for a model enzyme using RosettaDesign was reported. See: (a) Korkegian, A.; Black, M. E.; Baker, D.; Stoddard, B. L. *Science* **2005**, *308*, 857. Recently, a variant of the small protein RNase Sa (98 residues vs the ~ 260 residues of β -lactamases) was successfully prepared with a 28°C higher T_m than the wild-type protein. See: (b) Fu, H.; Grimsley, G.; Scholtz, J. M.; Pace, C. N. *Protein Sci.* **2010**, *19*, 1044. However, this large stabilization resulted from a combination of several stabilizing mutations previously identified using different rational approaches over a period of ~ 10 years, which actually emphasizes the difficulty of achieving T_m enhancements of several tens of degrees by rational design.
- (15) Bar-Even, A.; Noor, E.; Savir, Y.; Liebermeister, W.; Davivi, D.; Tawfik, D. S.; Milo, R. *Biochemistry* **2011**, *50*, 4402.
- (16) Levitt, P. S.; Papp-Wallace, K. M.; Taracila, M. A.; Hujer, A. M.; Winkler, M. L.; Smith, K. M.; Xu, Y.; Harris, M. E.; Bonomo, R. A. *J. Biol. Chem.* **2012**, *287*, 31783.
- (17) For instance, the mutation E104K is often found in clinical isolates, while E104 in TEM-1 β -lactamase is changed (to D, not K) in the highly promiscuous GNCA and PNCA lactamases. However, it does not appear likely that the E104D mutation by itself is responsible for the broad substrate scope, as there is also a D at position 104 in the GPBCA lactamase, which is much less promiscuous (Figure 3B).
- (18) Clarkson, M. W.; Lee, A. L. *Biochemistry* **2004**, *43*, 12448.
- (19) Glembo, T. J.; Farrell, D. W.; Gere, Z. N.; Thorpe, M. F.; Ozkan, S. B. *PLoS Comput. Biol.* **2012**, *8*, No. e1002428.
- (20) Villali, J.; Kern, D. *Curr. Opin. Chem. Biol.* **2010**, *14*, 636.
- (21) Indeed, given the significant difference in the sizes of the antibiotics studied (Figure S11), increased flexibility in functionally relevant regions of the promiscuous GNCA structure would seem plausible. Preliminary computational and experimental analyses (work in progress) point to enhanced flexibility in the catalytic α domain and enhanced rigidity in the $\alpha+\beta$ domain, a scenario that could rationalize the simultaneous occurrence of hyperstability and extended substrate scope. We note that the relation between conformational rigidity and stability, although plausible, does not necessarily hold in all cases. See: Jaenicke, R. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 2962.
- (22) The correlation between MIC and catalytic efficiency is not perfect (e.g., the increase in MIC for CTX upon “traveling back in time” 2–3 Gyr is smaller than the corresponding increase in catalytic efficiency), possibly reflecting molecular requirements related to export to the periplasm and subsequent folding of the lactamase enzymes. See: Chatzi, K. E.; Sardis, M. F.; Karamanou, S.; Economou, A. *Biochem. J.* **2013**, *449*, 25. For instance, export through the bacterial Sec system likely requires the preproteins to remain partially unfolded prior to interaction with the Sec channel, and this partial unfolding could be hindered by high stability.
- (23) Jensen, R. A. *Annu. Rev. Biochem.* **1976**, *30*, 409.
- (24) (a) Khersonsky, O.; Kiss, G.; Röthlisberger, D.; Dym, O.; Albeck, S.; Houk, K. N.; Baker, D.; Tawfik, D. S. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 10358. (b) Nobeli, I.; Favia, A. D.; Thornton, J. M. *Nat. Biotechnol.* **2009**, *27*, 157.

Chemical Biology Article Rankings: Current Top 10 – F1000

[Rankings/topten/current/faculty/chembiol](#)

Portal	GT Online Training	Course Surveys	GT Transcripts	TechWorks	Web Stats	GT Library	TechWorks	CoS Finance	CayuseGrants	New
------------------------	------------------------------------	--------------------------------	--------------------------------	---------------------------	---------------------------	----------------------------	---------------------------	-----------------------------	------------------------------	---------------------

Article Recommendations
MyF1000 [Sign out](#) [Subscribe](#)

ARTICLE RANKINGS

All Time Top 10
Current Top 10
All Time Most Viewed
Current Most Viewed
[Hidden Jewels](#)

BROWSE BY FACULTY

All | [Biology](#) | [Medicine](#)

- Anesthesiology & Pain Management
- Biochemistry
- Bioinformatics & Computational Biology
- Biotechnology
- Cancer Biology
- Cardiovascular Biology
- Cardiovascular Disorders
- Cell Biology
- Chemical Biology**
- Critical Care & Emergency Medicine
- Dermatology
- Developmental Biology
- Diabetes & Endocrinology
- Ecology
- Evolutionary Biology
- Gastroenterology & Hepatology
- Gastrointestinal Biology
- Genomics & Genetics
- Hematology
- Immunology
- Infectious Diseases
- Metabolic & Endocrine Science
- Microbiology
- Molecular Biology
- Molecular Medicine
- Nephrology
- Neurological Disorders
- Neuroscience
- Oncology
- Ophthalmology
- Otolaryngology
- Pharmacology & Drug Discovery
- Physiology
- Plant Biology
- Psychiatry
- Public Health & Epidemiology
- Renal Biology
- Research Methodology

F1000Prime » Rankings » Articles: Current Top 10 »

Chemical Biology

Current Top 10 rankings are generated every day using only the articles recommended in F1000Prime during the preceding 14 days. Articles are ranked according to their total star rating, which is calculated from all the stars awarded to that article by the Faculty. If any articles have the same total rating, then the most recently added article is listed first.

☐ [Save to MyF1000/Follow](#) [Export](#)

12

★

Principles for designing ideal protein structures.

Koga N, Tatsumi-Koga R, Liu G, Xiao R, Acton TB, Montelione GT, Baker D. Nature. 2012 Nov 8; 491(7423):222-7

[Abstract](#) | [Related Article Recommendations](#)

6 Recommendations

Charlotte Deane and Jamie Hill 27 Feb 2013 | Zelda R Wasserman | Reinhard Sterner and Rainer Merkl | Colin Kleanthous and Nick Housden | Xiayang Qiu and Ye Che | Nikolay Dokholyan and David Shrivanyants

[Good for Teaching](#)
[Interesting Hypothesis](#)
[New Finding](#)
[Technical Advance](#)

9

★

Genomic responses in mouse models poorly mimic human inflammatory diseases.

Seok J, Warren HS, Cuenca AG, Mindrinos MN, ..., Davis RW, Xiao W, Tompkins RG, Inflammation and Host Response to Injury, Large Scale Collaborative Research Program. Proc Natl Acad Sci U S A. 2013 Feb 26; 110(9):3507-12

[Abstract](#) | [Related Article Recommendations](#)

4 Recommendations

Linda Rinaman NEW 05 Mar 2013 | Luc Teyton | John Lowe | Mike Klimkowski

[Controversial](#)
[Good for Teaching](#)
[New Finding](#)

3

★

Inside the mind of a medicinal chemist: The role of human bias in compound prioritization during drug discovery.

Kutchukian PS, Vasilyeva NY, Xu J, Lindvall MK, Dillon MP, Glick M, Coley JD, Broojmans N. PLoS ONE. 2012; 7(11):e48476

[Abstract](#) | [Related Article Recommendations](#)

2 Recommendations

Jürgen Bajorath NEW 28 Feb 2013 | Martin Stahl

[Confirmation](#)
[Good for Teaching](#)
[New Finding](#)

3

★

Late-Stage Deoxyfluorination of Alcohols with PhenoFluor.

Sladojevich F, Arlow SI, Tang P, Ritter T. J Am Chem Soc. 2013 Feb 20; 135(7):2470-3

[Abstract](#) | [Related Article Recommendations](#)

1 Recommendation

Justin Chalker 26 Feb 2013

[New Finding](#)
[Technical Advance](#)

2

★

Hyperstability and substrate promiscuity in laboratory resurrections of Precambrian β-lactamases.

Risso VA, Gavira JA, Mejia-Carmona DF, Gaucher EA, Sanchez-Ruiz JM. J Am Chem Soc. 2013 Feb 27; 135(8):2899-902

[Abstract](#) | [Related Article Recommendations](#)

1 Recommendation

Justin Chalker NEW 01 Mar 2013

[Confirmation](#)
[Interesting Hypothesis](#)
[New Finding](#)
[Technical Advance](#)

F1000Research
 The fastest way to publish
[SUBMIT NOW](#)

COMING SOON:
NEXT-GEN PROTEIN PURIFICATION THAT SCALES TO FIT

[▶ Get the Guide](#)

ADVERTISEMENT

The F1000.com website uses cookies. By continuing to browse the site, you are agreeing to our use of cookies. [Find out more »](#)



Precambrian antibiotic resistance



Valeria A. Risso^{1,*}, Jose A. Gavira², Diego F. Mejia-Carmona¹, Eric A. Gaucher³ & Jose M. Sanchez-Ruiz^{1,*}

¹Facultad de Ciencias, Departamento de Química Física, Universidad de Granada, 18071 Granada, Spain.

²Laboratorio de Estudios Cristalográficos, Instituto Andaluz de Ciencias de la Tierra (CSIC – Universidad de Granada), 18100 Armilla, Granada, Spain.

³Georgia Institute of Technology, School of Biology, School of Chemistry, and Parker H. Petit Institute for Bioengineering and Biosciences, 30332 Atlanta, Georgia, USA.

INTRODUCTION AND OBJECTIVE

Experimental resurrection of ancestral proteins [1-2] provides insight into the ancient properties of biomolecules as well as intracellular and extracellular environments hosting these proteins. β -lactamases enzymes are widely distributed throughout the bacterial domain of life and are believed to have originated billions of years (Gyr) ago [3]. Here, we report the resurrection of Precambrian β -lactamases corresponding to several nodes in the phylogeny of Gram-negative bacteria. The ancient enzymes display high temperature stability, having denaturation temperature enhancements up to about 35 degrees Celsius with respect to modern β -lactamases. The most ancient of these β -lactamases are promiscuous towards a variety of substrates and are able to endow modern microorganisms with resistance towards different types of antibiotics (including third-generation). This enhanced promiscuity is not accompanied by significant changes in the active-site region as seen in static X-ray structures. This work supports the findings that Precambrian resurrections generate hyperstable proteins, provides direct evidence for the generalist-to-specialist conversion during protein evolution and opens up new possibilities in the study of the emergence of antibiotic resistance [4].

RESULTS

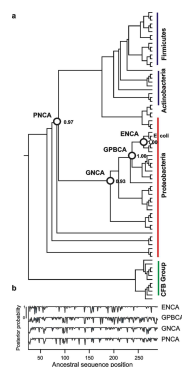


Figure 1. Phylogenetic analysis of β -lactamases and ancestral sequence reconstruction. (a) Cladogram of the tree used as a basis for the re-creation of Precambrian β -lactamases. The internal nodes targeted are labeled and their posterior probabilities shown. (b) Posterior probability profiles of the inferred amino acids across the 262 positions for the targeted internal nodes.

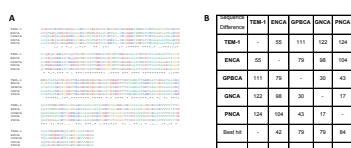


Figure 2. Sequence of TEM-1 β -lactamase and reconstructed sequences for the targeted Precambrian nodes. (A) The reconstructed sequences shown are the most probabilistic from each node (i.e., the sequences with the most probable amino acid at each position). (B) Sequence differences between the β -lactamases studied in this work.

CONCLUSION

- The resurrected proteins have substantial amino acid sequence differences with extant β -lactamases.
- These ancient β -lactamases can be described as generalist (promiscuous) enzymes, able to confer a moderate level of resistance against many different types of antibiotics.
- Crystallization and 3D-structure determination shows that the ancient lactamase shares the extant lactamase fold, despite extensive sequence differences (on the order of one hundred amino acid differences).
- These ancient enzyme lactamases are highly stable as demonstrated by Tm enhancements with respect to extant lactamases of about 35 degrees Celsius.

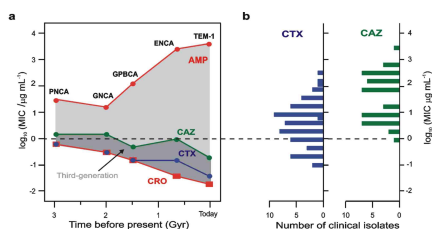


Figure 3. Functional characterization of resurrected Precambrian β -lactamases. (a) Minimum inhibitory concentration (MIC) values determined using an extant *E. coli* strain transformed with plasmids containing the Precambrian β -lactamases. Shading is meant to provide a visual illustration of the change of the range of values with geological time. Uncertainties in geological time are given in Fig 3a. (b) MIC values for third-generation antibiotics resulting from their clinical use (<http://www.lahey.org/Studies/temtable.asp>).

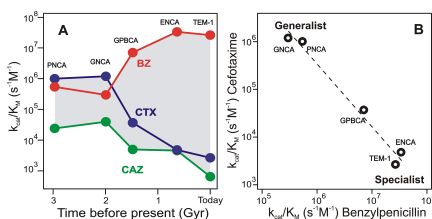


Figure 4. Functional characterization of laboratory resurrections of Precambrian β -lactamases. (a) Second-order rate constant for the in vitro hydrolysis of antibiotics catalyzed by extant and resurrected β -lactamases. (b) Generalist (broad substrate scope) to specialist conversion as illustrated by the catalytic efficiencies for benzylpenicillin and cefotaxime.

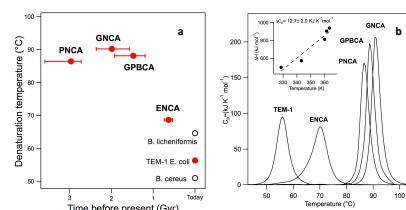


Figure 5. Hyperstability of resurrected Precambrian β -lactamases. (a) Denaturation temperature vs. geological time for β -lactamases. (b) The differential scanning calorimetry (DSC) profiles were obtained at a scan-rate of 200 K/hour and a protein concentration of 1 mg/mL.

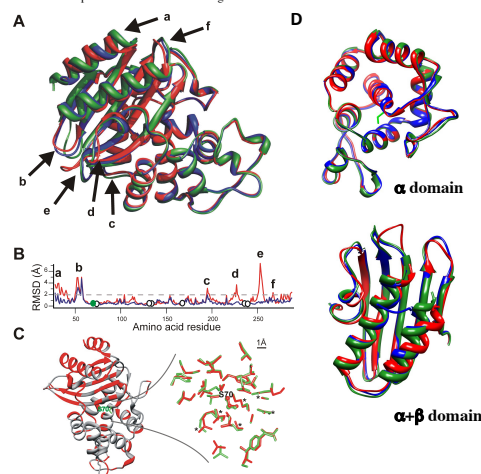


Figure 6. Structural characterization of laboratory resurrections of Precambrian β -lactamases. (a) Structural comparison of the TEM-1 β -lactamase (PDB.ID 1ht; green) and the resurrected β -lactamases for the last common ancestor of various Gram-negative bacteria (GNCA; PDB.ID 4b88; red) and the last common ancestor of enterobacteria (ENCA PDB.ID 3zdj; blue). The catalytic serine is shown. (b) RMSD plot; circles represent the active site positions (green for the catalytic serine). (c) GNCA lactamase structure (left) showing the positions with amino acid differences with respect to the modern TEM-1 lactamase. The conformation of the side-chains within a sphere of 8 Å centered at the catalytic serine 70 is also shown (right) for the GNCA (A) and TEM-1 (red and green). Active-site residues are labeled. (d) Structural comparison of the α and $\alpha\beta$ domains of TEM-1 β -lactamase from *E. coli* (green), GNCA (red) and ENCA (blue).

[1] Benner, S.A., Sassi, S.O. & Gaucher, E.A. (2007) Molecular paleoscience: systems biology from the past. *Adv. Enzymol. Relat. Areas Mol. Biol.* 75, 1-132.

[2] Perez-Jimenez, R. et al. (2011). Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nat. Struct. Mol. Biol.* 18, 592-596.

[3] Hall, B.G. & Barlow, M. (2004) Evolution of the serine β -lactamases: past, present and future. *Drug Resistance Updates* 7, 111-123.

[4] Risso VA, Gavira JA, Mejia-Carmona DF, Gaucher EA, Sanchez-Ruiz JM. (2013) Hyperstability and substrate promiscuity in laboratory resurrections of Precambrian β -lactamases. *JACS* 135, 2899-902.

Luis F. Castillo • Marco Cristancho
Gustavo Isaza • Andrés Pinzón
Juan Manuel Corchado Rodríguez
Editors

Advances in Computational Biology

Proceedings of the 2nd Colombian
Congress on Computational Biology
and Bioinformatics (CCBCOL)

FS-Tree: Sequential Association Rules and First Applications to Protein Secondary Structure Analysis

Nilson Mossos¹, Diego Fernando Mejia-Carmona², and Irene Tischer¹

¹ School of System Engineering and Computer Science, Universidad del Valle

² Graduate School of Biomedical Sciences, Universidad del Valle,
Calle 13 # 100 – 00, Cali, Colombia

Abstract. A protein's structure is determined by its amino acid sequence alone. In order to describe the relation between amino acid and corresponding structural sequences, we use an association rule mining approach. Traditional association rule mining is not appropriate in a sequential context. Therefore, we develop the structure FS-tree to represent subsequences and their frequencies in a sequence database, as well as the underlying construction algorithm.

A FS-tree is a prefix tree, which stores subsequences in a compact way. The sequential context oblige us to introduce a modification of the support concept, the relative support which does not give too much weight to short sequences. A 2-dimensional FS-tree for sequence pairs over different alphabets allows to obtain rules that establish the relation within the pairs.

Mining a 2-dimensional FS-tree of amino acid sequences and corresponding secondary structures, enables us to generate rules for their relation. We analyze hypothetical and observed tree size, inferring that there are short residue sequences acting as determinants of specific secondary structures. The most important rules are related to pure structure sequences, where rules for turn and helices exceed by far the rules for strands, as revealed by a rule composition analysis. By cross validation we verified that residue sequences with high propensity to specific structure sequences apply generally, independent of a specific protein sample. These promising results motivate us to explore FS-tree related analysis in a wider range of applications including the development of rules based prediction algorithms.

Keywords: algorithms, association rules, data mining, prediction, protein, secondary structure, propensity.

1 Introduction

Increasingly, data mining approaches are used to analyze and predict protein structure and function as well as interactions between proteins. Protein secondary structure analysis and prediction were initially studied by Chou and

Fasman [1,2] and GOR [3,4], who applied mainly statistical methods to determine the propensity of a group of 1-3 amino acids for a specific secondary structure. The propensity approach was extended to longer sequences[5]. Following studies include nearest neighbor algorithms [6,7] and neural networks[8]. The neural network approach was recently modified [9,10], resulting in a significant increase in prediction accuracy. Support vector machines, frequently used in the protein interaction problem, are successfully applied to predict secondary structure [11,12,13,14]. In this paper we explore the potential of association rules in secondary structure analysis and prediction, a data mining technique barely used in this context.

Originally, association rules were determined for a transactional database, where every transaction consists of a transaction ID and an itemset [15,16]. An association rule $P \Rightarrow Q$ is a relation between itemsets P and Q , such that $P \cap Q = \emptyset$. The support of a rule $P \Rightarrow Q$ is the frequency of $P \cup Q$, its confidence represents the probability that Q occurs, whenever P occurs. A rule $P \Rightarrow Q$ is defined as interesting, if its support and confidence values exceed a chosen threshold. Rule extraction addresses the frequent itemset mining problem, which is generally resolved by the candidate-generate-and-test approach proposed by Agrawal [16] or by pattern growth introduced by Han[17]. Han employs a prefix tree, the Frequent Pattern Tree (FP-tree), to store itemsets, ordered by item frequency and eliminating not frequent items. The FP-tree is a compact representation of itemsets, because common prefixes share a path of nodes. The rule mining problem for biological sequence analysis has to deal with sequential, not transactional databases. We need to identify frequent subsequences, being most important the order in which symbols occur. Therefore, we create the data structure Frequent Sequence Tree (FS-tree), capable to efficiently store subsequences occurring in a sequential database together with their frequencies. We develop the algorithms for FS-tree construction and frequent subsequences extraction. Applying the FS-tree construction and extraction to paired sequence data (P, Q) , where P and Q are sequences of different types, we are able to generate association rules of the form $P \Rightarrow Q$ that establish the relation between P and Q quantified by the corresponding support and confidence values.

Constructing the 2-dimensional FS-tree for data pairs of proteins and their secondary structure sequences we are able to extract association rules in order to determine those amino acid subsequences, that assume a specific structural sequence with high frequency. We explore the bioinformatic potential of this FS-tree, analyzing the obtained tree size and the resulting association rules, and we show that the most interesting rules do not depend on the specific data set used to generate the FS-tree.

The rest of this document is organized as follows: Section 2 contains the informatic development: tree construction and sequence extraction; section 3 applies the FS-tree to protein structure analysis; and finally, section 4 is dedicated to our conclusions.

2 Tree Construction and Sequence Extraction

2.1 FS-Tree Construction

Algorithm `buildFS-tree( $D$ )`
 Input: sequence database D
 Output: its frequent sequence tree T ; level frequency array L
 Create the root of an FS-tree T ,
 and label it as "null"
 Initialize L in 0
 for all sequences $X \in D$
 $S_X = \text{set of suffixes of } X$
 for all $s \in S_X$
 `increasePath( $T, L, s$ )`

Fig. 1. Algorithm for tree construction

The Frequent Sequence Tree (FS-tree) we propose here, stores efficiently all subsequences from a sequence database D over a finite alphabet Σ , along with their frequencies. The tree is constructed from the suffixes of all sequences in D . Every suffix is represented as path obtained by traversing the tree in depth from its root. Each node N stores a symbol from alphabet Σ and a frequency count. The shortest path between root and N represents a suffix of a subsequence in D , the count establishes its frequency in D . Sequences in D with common suffix s in the database share this

suffix in the tree, allowing to refer to a suffix as s , whether it is subsequence in the database or path in the tree. The resulting tree is able to store all subsequences occurring in D . It is compact because common suffixes are shared. The stored suffix frequencies allow easy extraction of frequent subsequences. The algorithm responsible for the tree construction, `buildFS-tree` is shown in figure 1.

2.2 Frequent Sequence Extraction

In the current situation, the concept of support expressed directly in frequency seems to be not the most appropriate one. The shorter a sequence, the higher is its probability to appear in the database and this high probability could shadow the relative frequent appearance of interesting longer sequences. Therefore, we introduce the concept of relative support of a sequence s as its frequency with respect to the set of all sequences of the same length: A sequence is defined to be frequent, if its relative support exceeds a given support threshold *minsup*. Note that different to the traditional support concept, the relative support does not necessarily decrease with the tree level.

The level frequency L updated during tree construction, is used to extract frequent sequences. The iterative extraction algorithm applies an in-depth search. Due to the inequality

$$\text{relative support}(h) \leq \frac{\text{frequency in } h}{L_{\text{level}(h)}} \leq \frac{\text{frequency in } c}{L_{\text{height}}}$$

it is possible to prune the search process at a descendant h of a node c whenever

$$\frac{\text{frequency in } c}{L_{\text{height}}} \leq \text{minsup}.$$

2.3 Time Analysis and Space Requirements

Let D be a database of n sequences over an alphabet Σ of size s and m the maximum length of sequences in D .

Algorithm **buildFP – tree** is analyzed in the worst-case scenario, where all database sequences have length m . For each sequence in D , a set of m suffixes will be generated, with lengths ranging from 1 to m , requiring an execution cost of $\mathcal{O}(m)$. The cost of sending a suffix to **increasePath** is proportional to its length; therefore, we obtain a total cost of $\mathcal{O}(\sum_{j=1}^m j) = \mathcal{O}(m^2)$ for constructing the paths for one sequence; considering all n sequences in D , the resulting worst-case cost for **buildFP – tree** is $\mathcal{O}(n \cdot m^2)$. The worst-case for space requirements is $\mathcal{O}(m^s)$, according to the size of a complete s -ary tree of depth m .

The frequent subsequence extraction is an in-place algorithm, hence no additional space is required. The execution time, given by the number of visited nodes, is $\mathcal{O}(m^s)$.

2.4 Frequent Sequence Pairs and Rule Extraction

According to our bioinformatics goals, we need to construct a FS-tree for sequence pairs and extract its interesting rules, i.e. given is a database D of sequence pairs (P, Q) , where P and Q have same length and are defined over the finite alphabets Σ_1 and Σ_2 , respectively. The problem is reduced to the previous case, considering the 2-dimensional alphabet $\Sigma = \Sigma_1 \times \Sigma_2$ and the transposed 2-dimensional sequences $Z = (P, Q)^t$. Subsequence pairs are frequent (section 2.2) if a user defined relative support threshold *minsup* is exceeded. A rule is a frequent subsequence pair (P, Q) for which also a confidence threshold *minconf* holds. The confidence of a frequent pair depends on the frequency in which P occurs as antecedent. In order to obtain the confidence of a frequent pair we have to search the tree in depth first and totalize the frequencies of nodes with antecedent P , extracting rules only if *minconf* is exceeded.

3 First Applications to Protein Structure Analysis

A main problem in protein structure analysis is to determine the tendency of an amino acid sequence P for a secondary structure sequence Q . Our approach applies FS-tree construction and rule extraction, identifying rules $P \Rightarrow Q$, which exceed given thresholds for support and confidence.

3.1 Data Set and Tree Construction

The proposed mining technique is based on data from Protein Data Bank (PDB) which offers high quality and quantity information of protein sequences together with their 3-dimensional atom coordinates[18]. We based our analyses on a subset PDB, referred as PDBsub, that consists of proteins solved by X-ray diffraction, removing sequences with non-standard amino acid residues. PDBsub contains

39.660 protein sequences, composed of 11'039.951 residues, 37,82 % of them in helix, 22,27% in strand and 39.92% in turn. To obtain the secondary structure sequences, we applied the DSSP algorithm [19] an algorithm that assigns the secondary structure from the atomic coordinates of a protein, based on hydrogen bond pattern detection; grouping in one 3 types of helix (H, G, I) labeled as H, 2 strands (B, E) labeled as B and 3 coils (T, S, U) labeled as U. We constructed the FS-tree of pairs, as described in section 2.4, using the alphabet Σ_1 of amino acids (1-letter code) and the alphabet of secondary structures $\Sigma_2 = \{H,B,U\}$.

3.2 Real versus Hypothetical FS-Tree Size for PDBsub

At each tree level k , the hypothetical number of nodes is given by $(20 \times 3)^k = 60^k$, according to the possibility of combining in a sequence of length k , 20 residues with 3 structures. Contrasting the number of hypothetical nodes with the observed ones we obtain table 1.

Table 1. Observed and hypothetical FS-tree size

level k	1	2	3	4	5
subsequences of length k	11'039.951	10'978.935	10'918.836	10'859.045	10'799.491
observed nodes	60	3.538	119.641	1'176.938	2'914.686
hypothetical nodes	60	3.600	216.000	12'960.000	777'600.000

At level 1, all 60 theoretically possible nodes are found in our sample set, which is not surprising, given that our sample contains over 11 million residues in different secondary structures. That means, that every residue occurs in each, helix, strand and turn. The frequency in each structure is shown in figure 2. We observe that Proline and Glycine are the amino acids with the highest frequencies in a specific secondary structure; appearing over three times more in turns than in others. Valine is the only amino acid with the highest frequency for strands. Alanine, Glutamic acid, and Leucine are the amino acids with a dominant frequency for helices. Cysteine, Phenylalanine, Isoleucine, Lysine, Valine and Tyrosine are amino acids whose frequency is not dominant in any secondary structure. Given the frequencies of individual amino acids at the FS-tree level 1, it is easy to calculate the classic Chou and Fasman propensities of amino acids, for any given data set and extent the propensity concept to longer sequences (see section 3.3). This could be an interesting application of the FS-tree structure, especially if we keep in mind that recent publications ([20,21]) state that propensity depends on the selected protein data set. We calculated the propensities of individual amino acids based on the FS-tree and found nearly perfect coincidence between our results and the propensities determined for PDBselect in [21]. At level 2, not all possible nodes are present. Table 2 shows the residue-structure pairs of length 2 which are not observed in our dataset. Each residue pair occurs in any of the considered pure structures (HH, BB, UU) - obviously with very different frequencies. But there are some residue pairs, that are not

found in the transitional structures BU, HU, BH or HB. Proline seems to play a special role in pure structures, as many possible rules relating Proline with a transitional structure are not observed.

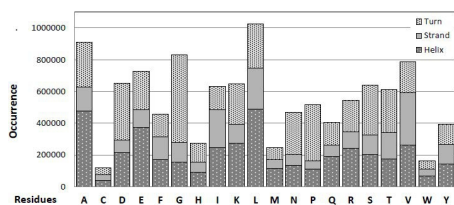


Fig. 2. Analysis of tree level 1: Frequency of residues in secondary structures

Table 2. Analysis of tree level 2: Residue-structures pairs not present in PDBsub

Residue sequence	not combined with	structure sequence
PP.		BU
CP, EP, NP, PH, PI, PR, PV, PY, WP, YP.		HU
MF, PP, QC, WC.		BH
AP, CD, CP, DP, EP, EQ, FP, GP, HP, HW, IP, IW, KP, LP, MP, NP, PA, PC, PD, PE, PF, PG, PH, PI, PK, PL, PM, PN, PP, PQ, PR, PS, PT, PV, PW, PY, QP, RP, SP, TP, VH, VP, WN, WP, WS, WW, YP.		HB

At level 3, only 119.641 out of 216.000 possible nodes occur (55.39%). More than 10 million residue-structure triplets are distributed over the observed nodes, corresponding to an average frequency of 91.26 per residue-structure triplet. At level 4 and higher, the number of observed nodes does not reach the hypothetic quantity, *i.e.* our sample set PDBsub and even PDB (PDBsub is roughly half the size of PDB) are too small to conclude whether all combinations naturally exist. Nevertheless, the node frequency at each level is relatively high (about 3 at levels 5-8), and there are nodes significantly more frequent than others (rules), which suggests that there are short residue sequences acting as determinants of specific secondary structures.

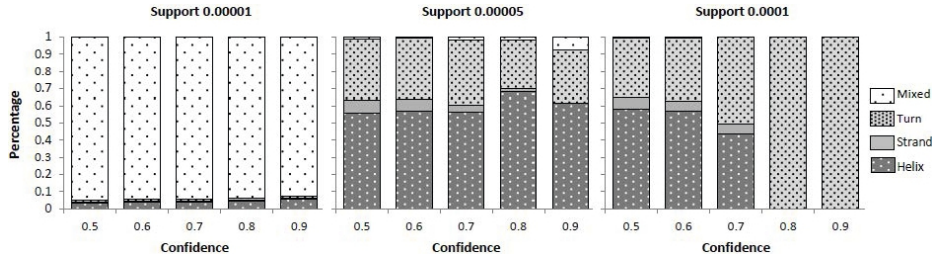
3.3 Analysis of Rule Number and Composition

A rule's support refers to its frequency in the dataset, its confidence indicates the preference of the residue sequence to a specific structural sequence. Together, support and confidence allow to identify relations between residual and structural sequences which are very probable. These interesting rules, easily extracted from the FS-tree, could be considered as an extension of the propensity concept for longer sequences. Different to the classic propensity, this extended concept allows to adjust the desired grade of certainty of selected rules defining adequate support and confidence thresholds. For instance, in the case of sequences of length 1, using a threshold of 10^{-6} for support and 0.41 for confidence, we extracted 16 interesting rules, all of them with a classic propensity greater than 1.1; for thresholds of 10^{-5} and 0.445 respectively, we obtain only 9 rules, all of them with a classic propensity greater than 1.3.

We analyzed number and composition of interesting rules, in dependence of support and confidence thresholds. As expected, the number of rules decreases as support and confidence increase (see table 3 for some specific values).

Table 3. Rules in dependence of support and confidence

Support	Confidence				
	50%	60%	70%	80%	90%
10^{-5}	102.934	94.099	84.447	67.647	51.205
$5 \cdot 10^{-5}$	1.264	598	187	57	13
10^{-4}	648	316	67	5	1

**Fig. 3.** Rule composition according to secondary structure**Table 4.** Rules for thresholds of 0.00005 (support) and 0.8 (confidence)

Residue sequence	structure sequence
PP.	UU
FQM.	HHH
PPP, QPD, PPG, PDG, KPP, GPP, QQP, DPP, NPP, EPP, HPN.	UUU
TVLV.	BBBB
EQVL, LQKV, KLFN, RALA, LAKE, RNLL, ELDK, DEAE, RAKR, DEAA, KSEL, EAEL, EMLR, NAAK, AAKS, DKAI, DLLE, EKLF, EALA, ELAR, VEAA, EEAL, AVRR, DAAV, AAVN, EELL, EALR, LDKA, QKV V, EAAR, EAEK, RAAL, AELL, KALE, AAVR, EAAV, VDAA.	HHHH
RDLK, QPDG, HRDL.	UUUU
VDAAV.	HHHHH
RDLKP.	UUUUH
HRDLK.	UUUUU

In figure 3 we show the rule composition for specific support and confidence values. We observe that the percentage of mixed rules -rules that involve more than one structure type- decreases as support and confidence are increased. Accordingly, the percentage of pure rules -where only one structure is present- increases for higher support and confidence thresholds: the most important rules predict pure structures; *i.e.* certain residue sequences occur frequently in a

specific secondary structure type. These sequences could be considered as elementary units for secondary structure determination, or, in other words, sequences with a very high propensity to a specific secondary structure.

The rules obtained for thresholds of 0.00005 (support) and 0.8 (confidence) are shown in figure 4. The most interesting rules represent a “pure”, not mixed secondary structure (with only one exception, the rule $\text{RDLKP} \Rightarrow \text{UUUUH}$). All are short rules, the maximum length is 5. The shortest rule is $\text{PP} \Rightarrow \text{UU}$; *i.e.* a sequence of 2 Prolines is most likely to be found in a turn. Among the 12 third-level rules that exceed the chosen thresholds there is only one rule for a helix structure ($\text{FQM} \Rightarrow \text{HHH}$). The others are rules for turns. In all rules that derive UUU, there is at least one Proline in the corresponding residue sequence (4 sequences with one Proline, and 7 with 2 or more Prolines). This is coherent with the findings of Costantini *et al.* [21], who offered an update of the Chou and Fasman propensities, by using the PDBselect, with more than 2000 proteins: they suggest that Proline is the most determinant amino acid for a secondary structure, with a propensity for turns, that is over three times higher than for any other structure. Accordingly, our strongest rule (support 0.00013, confidence 90%) is $\text{PPP} \Rightarrow \text{UUU}$; *i.e.* PPP is most probably found in turns. This supports the suggestion to use Prolines to increase stability of turns ([22]). At level 4, the dominant structure is helix, present in 37 of 41 rules. Almost all residue sequences for helices contain at least one of the amino acids with the highest propensities for helices (Alanine (A), Glutamic Acid (E) and Leucine (L)); being $\text{QKVV} \Rightarrow \text{HHHH}$ the only exception. At this level we also extract 3 rules for turns, and the only interesting rule for strands $\text{TVLV} \Rightarrow \text{BBBB}$. At level 5, there are only 3 rules, each one continues a rule of length 4: $\text{VDAAV} \Rightarrow \text{HHHHH}$, $\text{RDLKP} \Rightarrow \text{UUUUH}$ and $\text{HRDLK} \Rightarrow \text{UUUUU}$. It is curious that in fifth-level rules, the HHHHH and UUUUU do not contain the most determinant residues for those secondary structures (A, E, L for H and P for U).

3.4 Domain of Interesting Rules

Previously we derived the interesting rules for chosen support and confidence thresholds. In which degree do these rules depend on the specific data set, used to construct the FS-tree? To answer this question we applied a seven-fold cross-validation, dividing PDBsub and its corresponding structure sequences in 7 parts. In each validation process, 6 parts are used for training; the seventh part is reserved to test if the rules apply. We used the training data to construct the FS-tree and to capture the rules. Then we verified in the test set the rule application: for each incident of a rule’s residue sequence we verified if it appears with the corresponding structural sequence. The percentage of rules that apply correctly, is reported in table 5.

The table shows that the most important rules (confidence 90%) apply in a very high percentage in the test sets, suggesting that these rules do not depend on the data set that is used to construct the tree. Therefore we can conclude that the residue sequences with high propensity to specific structure sequences apply generally, not only in a specific context.

Table 5. Percentage of rules generated by the training set, which apply in the test set

Support	Confidence				
	50%	60%	70%	80%	90%
10^{-5}	74.269±0.011	83.142±0.012	89.295±0.011	93.153±0.012	96.835±0.011
$5 \cdot 10^{-5}$	58.527±0.002	66.524±0.002	74.666±0.005	84.671±0.011	93.037±0.010
10^{-4}	58.405±0.002	66.299±0.002	73.431±0.005	82.410±0.009	89.978±0.025

4 Conclusions

We present the FS-tree, a new prefix tree structure, capable to store compactly subsequences of a sequential database. In a FS-tree, a pattern is an ordered sequence of symbols, whereas the traditional FP-tree for transactional databases is based on sets. For sequential databases, the standard concept of support has to be revised. In this context it is important to determine the frequency of a sequence with respect to sequences of equal length, which leads us to the definition of relative support. We define an efficient algorithm for extracting frequent sequences, based on the appropriate strategy to reduce FS-tree search. Applying the FS-tree structure and algorithms to sequence pairs, we are able to generate sequential association rules between sequences of two different alphabets. In order to illustrate the potential of FS-trees in bioinformatics, we apply it to secondary structure analysis of protein sequences.

Our first analysis concerns the number of observed nodes of fixed length. As an overall result we obtain, that neither all hypothetically possible short residue sequences nor all possible structure sequences for a given residue sequence are observed. A possible explanation of these results includes both, the limited number of structurally known proteins as well as the propensity of residue sequences for specific secondary structures. Secondly we analyzed the number and composition of interesting rules, in dependence of the chosen support and confidence thresholds. The most important rules are related to pure structure sequences, where rules for turn and helices exceed by far the rules for strands. The obtained rules provide the residue-structure relations with the highest propensity. By cross validation we showed that the most frequent rules are general, not depending on the specific data set used to generate the tree.

These first analyses show the high potential of the FS-tree structure in bioinformatics and motivate us to explore it in a wider range of applications, which include the development of rules based prediction algorithms.

References

1. Chou, P.Y., Fasman, G.D.: Prediction of protein conformation. *Biochemistry* 13(2), 222–245 (1974)
2. Chou, P.Y., Fasman, G.D.: Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry* 13(2), 211–222 (1974)

3. Garnier, J., Osguthorpe, D.J., Robson, B.: Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of Molecular Biology* 120(1), 97–120 (1978)
4. Garnier, J., Gibrat, J.F., Robson, B.: GOR method for predicting protein secondary structure from amino acid sequence. *Methods in Enzymology*, vol. 266, pp. 540–553. Academic Press (1996)
5. Strelets, V.B.: New machine learning technique for analysis and prediction of sequence and structure features: Protein secondary structure prediction. *Network Science* (1995)
6. Salzberg, S., Cost, S.: Predicting protein secondary structure with a nearest-neighbor algorithm. *Journal of Molecular Biology* 227(2), 371–374 (1992)
7. Salamov, A.A., Solovyev, V.V.: Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *Journal of Molecular Biology* 247(1), 11–15 (1995)
8. Rost, B.: Phd: predicting one-dimensional protein structure by profile-based neural networks. *Methods in Enzymology* 266, 525–539 (1996)
9. Pollastri, G., McLysaght, A.: Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics* 21(8), 1719–1720 (2005)
10. Zhang, G.Z., Huang, D.S., Zhu, Y.P., Li, Y.X.: Improving protein secondary structure prediction by using the residue conformational classes. *Pattern Recognition Letters* 26(15), 2346–2352 (2005)
11. Bock, J.R., Gough, D.A.: Predicting protein-protein interactions from primary structure. *Bioinformatics* 17(5), 455–460 (2001)
12. Bradford, J.R., Westhead, D.R.: Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics* 21(8), 1487–1494 (2005)
13. Ward, J.J., McGuffin, L.J., Buxton, B.F., Jones, D.T.: Secondary structure prediction with support vector machines. *Bioinformatics* 19(13), 1650–1655 (2003)
14. Birzele, F., Kramer, S.: A new representation for protein secondary structure prediction based on frequent patterns. *Bioinformatics* 22(22), 2628–2634 (2006)
15. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data SIGMOD 1993*, vol. 22, pp. 207–216 (May 1993)
16. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules in Large Databases, vol. 15, pp. 487–499. Morgan Kaufmann Publishers Inc. (1994)
17. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. *ACM SIGMOD Record* 29(2), 1–12 (2000)
18. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The protein data bank. *Nucleic Acids Research* 28(1), 235–242 (2000)
19. Kabsch, W., Sander, C.: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22(12), 2577–2637 (1983)
20. Fujiwara, K., Toda, H., Ikeguchi, M.: Dependence of alpha-helical and beta-sheet amino acid propensities on the overall protein fold type. *BMC Structural Biology* 12(1), 18 (2012)
21. Costantini, S., Colonna, G., Facchiano, A.M.: Amino acid propensities for secondary structures are influenced by the protein structural class. *Biochemical and Biophysical Research Communications* 342(2), 441–451 (2006)
22. Fu, H., Grimsley, G.R., Razvi, A., Scholtz, J.M., Pace, C.N.: Increasing protein stability by improving beta-turns. *Proteins* 77(3), 491–498 (2009)