UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

INSTITUTO DE ECOLOGÍA

Sistemática

Identificación de patrones de hidrofobicidad distintivos y la utilidad de la distribución de aminoácidos hidrofóbicos para inferir agrupamientos estructurales entre proteínas de secuencia divergente. El caso de las Lisozimas de Bacterias.

TESIS

QUE PARA OPTAR POR EL GRADO DE:

MAESTRO EN CIENCIAS BIOLÓGICAS

PRESENTA:

Octavio Abraham Moreno Guillén

TUTORA PRINCIPAL:

Dra. Luisa Isaura Falcón Álvarez, Instituto de Ecología, UNAM

COMITÉ TUTOR:

Dr. León Patricio Martínez Castilla, Facultad de Química, UNAM

Dr. Luis J. Delaye Arredondo, Posgrado en Ciencias Biológicas, UNAM

MÉXICO, D.F. Noviembre 2014

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

INSTITUTO DE ECOLOGÍA

Sistemática

Identificación de patrones de hidrofobicidad distintivos y la utilidad de la distribución de aminoácidos hidrofóbicos para inferir agrupamientos estructurales entre proteínas de secuencia divergente. El caso de las Lisozimas de Bacterias.

TESIS

QUE PARA OPTAR POR EL GRADO DE:

MAESTRO EN CIENCIAS BIOLÓGICAS

PRESENTA:

Octavio Abraham Moreno Guillén

TUTORA PRINCIPAL:

Dra. Luisa Isaura Falcón Álvarez, Instituto de Ecología, UNAM

COMITÉ TUTOR:

Dr. León Patricio Martínez Castilla, Facultad de Química, UNAM

Dr. Luis J. Delaye Arredondo, Posgrado en Ciencias Biológicas, UNAM

MÉXICO, D.F. Noviembre 2014

UNAM
POSGRADO
Ciencias Biológicas

COORDINACIÓN

Dr. Isidro Ávila Martínez
Director General de Administración Escolar, UNAM
P r e s e n t e

Me permito informar a usted, que el Subcomité de Biología Evolutiva y Sistemática, en su sesión ordinaria del día 29 de septiembre de 2014, aprobó el jurado para la presentación de su examen para obtener el grado de **MAESTRO EN CIENCIAS BIOLÓGICAS del Posgrado** en Ciencias Biológicas, del alumno **MORENO GUILLEN OCTAVIO ABRAHAM** con número de cuenta **300020074** con la tesis titulada: **"IDENTIFICACIÓN DE PATRONES DE HIDROFOBICIDAD DISTINTIVOS Y LA UTILIDAD DE LA DISTRIBUCIÓN DE AMINOÁCIDOS HIDROFÓBICOS PARA INFERIR AGRUPAMIENTOS ESTRUCTURALES ENTRE PROTEÍNAS DE SECUENCIA DIVERGENTE. EL CASO DE LAS LISOZIMAS DE BACTERIAS"**, bajo la dirección de la **DRA. LUISA ISAURA FALCÓN ÁLVAREZ:**

Presidente:	DR. LORENZO PATRICK SEGOVIA FORCELLA
Vocal:	DR. LUIS DAVID ALCARAZ PERAZA
Secretario:	DR. LEÓN PATRICIO MARTÍNEZ CASTILLA
Suplente:	DRA. MARIANA BENITEZ KEINRAD
Suplente:	DR. LUIS JOSÉ DELAYE ARREDONDO

Sin otro particular, me es grato enviarle un cordial saludo.

A T E N T A M E N T E
"POR MI RAZA HABLARA EL ESPIRITU"
Cd. Universitaria, D.F., a 19 de noviembre de 2014.

DRA. MARÍA DEL CORO ARIZMENDI ARRIAGA
COORDINADORA DEL PROGRAMA

POSGRADO EN CIENCIAS BIOLÓGICAS
COORDINACIÓN

c.c.p. Expediente del (la) interesado (a).

AGRADECIMIENTOS INSTITUCIONALES

> Al CONACyT, proyecto #151796, por la beca y facilidades otorgadas durante la realización de mis estudios de Maestría en el Posgrado en Ciencias Biológicas.

> A PAPIIT IT-100212-3 por el financiamiento otorgado al grupo de trabajo en donde se realizó este trabajo.

> Al Posgrado en Ciencias Biológicas, U.N.A.M., por permitirme desempeñarme académicamente bajo sus programas de estudio, así como por la oportunidad de participar en las ofertas académicas disponibles (seminarios, conferencias, pláticas, etc.)

> A los miembros de mi comité tutoral, la Dra. Luisa I. Falcón Álvarez, al Dr. León Patricio Martínez Castilla y al Dr. Luis J. Delaye Arredondo, por todo su apoyo, conocimiento, experiencia y guía otorgados para el mejoramiento de mi trabajo de investigación.

AGRADECIMIENTOS ACADEMICOS

> Al Laboratorio de Ecología Microbiana del Instituto de Ecología, U.N.A.M. por todas las facilidades inmobiliarias durante mis estudios de Maestría.

> Al Instituto de Ecología U.N.A.M., por todas las facilidades inmobiliarias para llevar a cabo mis estudios de Maestría, así como por la oferta académica de alto nivel (seminarios, conferencias, pláticas, etc.).

> A los miembros del sínodo, el Dr. Lorenzo Segovia, Dr. León P. Castilla, Dr. Luis Delaye, Dra. Mariana Benítez y al Dr. Luis David Alcaraz por todo su tiempo y esfuerzo para revisar mi trabajo lo más pronto posible. Agradezco sus comentarios y observaciones que me permitieron mejorar la calidad de mi trabajo.

> A los Técnicos Académicos, M. en C. Osiris Gaona Pineda y al M. en C. Antonio Cruz por su ayuda en laboratorio, así como en facilidades y aportaciones para la realización de mi trabajo.

AGRADECIMIENTOS PERSONALES

Son muchas las personas a las que le tengo que agradecer por su ayuda y/o apoyo durante la realización de mis estudios de Maestría. Si omito a alguien espero me perdonen.

> Primeramente, estoy en deuda con la Bióloga Verónica (RodoVero) Marlene Cruz Hernández por toda la ayuda con aspectos técnicos de esta tesis; de forma muy notoria, por la ayuda otorgada con el dendograma estructural. Esa figura existe gracias a sus consejos y habilidad en el diseño y edición de imágenes. Mi querida RodoVero, gracias por tu amistad, paciencia y cariño. Gracias por darme a mi pequeña hija, a quien amo y aprecio mucho.

> A mis padres, Irma Guillén y Antonio Moreno por toda su ayuda, apoyo, apoyo moral, alojo y demás durante toda mi vida (no sólo durante la maestría).

> A mi tutora, la Dra. Luisa I. Falcón Álvarez por su amistad y paciencia. Gracias Luisa, aprendí mucho de ti.

> Al Dr. Luis Delaye, a quien considero un gran amigo y profesor. Este agradecimiento es porque, aun con la distancia de por medio, tuvo la mejor disposición que cualquier alumno podría pedir para ayudarme con los trámites de titulación, la revisión de mi tesis, las reuniones de los tutorales, etc. Muchas gracias Luis!!!.

> Al burrito. Todas las veces que hemos ido al parque me han sido muy útiles para tomar decisiones técnicas que plegaron la realización de este trabajo de tesis. Espero todavía le pueda otorgar otro reconocimiento en la tesis de Doc. Burrito, tú aguanta aunque ya estés viejito, yo me apuro.

AGRADECIMIENTOS MUY ESPECIALES A:

> La Dra. Rocío Jetzabel (Evil Dra. Rooooo) Alcántara H., por toda su amistad, compañía, simpatía y por quedarse hasta tarde conmigo en los fines de semestre mientras yo terminaba todo mi trabajo atrasado. Agradezco ésas pláticas Mega-Nerds acerca modelos evolutivos, la incertidumbre de la distribución de Poisson, las conjeturas con los métodos de distancia y las paradojas de la vida.

> La Bióloga Mariana (Arenita) Argott, por su ayuda, paciencia, compañía y todas las sonrisas que siempre me sacó.

> La Bióloga Teresa (Ms. PCR) Pérez Carbajal, por tolerar todos mis ataques de neuras y por su compañía en los momentos difíciles. Gracias además por ayudar a mantener mi ego bajo control.

> La M. en C. Rocío V. (Maestraza Roooxxxiiiooo) González V., por su amistad y ayuda infinita con la tramitología.

> Al M. en C. Mario (Maestrazo Mario) Carrillo, por aguantarme tanto tiempo.

> También a todos los del labo, como Yislem (Dra. Grinchlem), Patricia Valdespino (Pa-Pa-Pancrónica), Daniel y Osiris, gracias a todos por su amistad y paciencia.

> Finalmente, a cualquier colado que sienta que contribuyó, pero a mi consideración no fue suficiente como para aparecer en esta honorable sección.

DEDICATORIA

Le dedico esta tesis (y todo el esfuerzo que me requirió concluirla) a mi hija, Emma H. Moreno Cruz, a quien quiero con todo mi amor, aprecio y respeto.

INDICE

1. Resumen.....1
1.1 Abstract..... 2
2. Introducción..... 3
2.1 Estructura de Proteínas..... 3
2.2 La importancia de los aminoácidos en la estructura y función de las proteínas..... 4
2.3 Consideraciones importantes en el análisis de la estructura tridimensional..... 7
2.4 Los aminoácidos hidrofóbicos como “caracteres estructurales” en
la comparación de proteínas.....12
2.5 El Método del Análisis de los Agrupamientos (*Clusters*) Hidrofóbicos (HCA).....14
2.6 Ventajas y desventajas del método de HCA..... 16
2.7 Los Diagramas de HCA.....19
2.8 Representaciones de los aminoácidos en los diagramas de HCA..... 24
3. Las Lisozimas.....27
4. Justificación..... 33
4.1 Hipótesis..... 35
4.2 Objetivos generales..... 35
4.3 Objetivos particulares.....35
5. METODOLOGIA.....36
5.1 El dendograma estructural..... 36
5.2 La Búsqueda Secuencias Homólogas Distantes de la Lisozima Tipo C.....37
5.3 Alineamiento de las Secuencias.....38
5.4 La Matriz de HCA..... 39
5.5 Parámetros importantes en el alineamiento.....40
5.6 El algoritmo HCA_Analyze_Multiple_Aligns (HCAmA).....41
5.7 Análisis de identidad entre los agrupamientos hidrofóbicos de las secuencias analizadas..... 43
5.8 Los Patrones Distintivos de Hidrofobicidad (PDH)..... 46
6. RESULTADOS..... 48
7. DISCUSION.....51
7.1 Las Lisozimas de Bacterias como modelo de análisis estructurales..... 51
7.2 Problemática con las Lisozimas de Bacterias.....54
7.3 Utilidad de los Patrones Distintivos de Hidrofobicidad (PDH).....57
7.4 Detalles importantes de las secuencias analizadas..... 59

7.5 Agrupamientos estructurales.....60
7.6 Representatividad de los PDH.....62
8. CONCLUSION.....68
9. PERSPECTIVAS.....72
10. GLOSARIO.....74
11. REFERENCIAS.....78
12. ANEXOS.....83

INDICE DE FIGURAS **Página #**

Figura 1. **Página 4**
Los cuatro niveles de organización estructural en las proteínas.

Figura 2. **Página 6**
Los veinte aminoácidos y su clasificación en polares, no-polares y anfipáticos.

Figura 3. **Página 20**
Interpretando un diagrama de HCA.

Figura 4. **Página 21**
Representación de la distancia de conectividad.

Figura 5. **Página 23**
Representación de una secuencia de aminoácidos en código binario para su análisis mediante el método de HCA.

Tabla 1. **Página 24**
Tabla de la frecuencia de ocurrencia de los aminoácidos dependiendo del tipo de estructura secundaria que tienden a formar.

Figura 6. **Página 25**
Representación mediante los diagramas de HCA de la estructura de la Lisozima tipo C de humano.

Tabla 2. **Página 26**
Características generales resumidas de los aminoácidos importantes en el método de HCA.

Figura 7. **Página 47**
Esquema de la metodología seguida.

Figura 8. **Página 50**

El Dendograma estructural.

Tabla 3. **Página 51**

Listado de los 49 Patrones Distintivos de Hidrofobicidad (PDH).

Figura 9. **Página 63**

La secuencia de *Tannerella sp.*, es el PDH más frecuente.

Figura 10. **Página 67**

Tabla de frecuencias de los 49 PDH.

1. Resumen

Con el explosivo aumento en el número de secuencias codificantes que provienen de los diversos genomas secuenciados, ha surgido el problema de asignar identidad a toda esa vasta cantidad de información que se acumula día con día. Una forma práctica de facilitar y acelerar la categorización de la "nueva" información es la comparación sistemática de las secuencias recién obtenidas con base en otras ya anotadas. Aunque en principio, ésta labor de comparación implica sólo la automatización del proceso, con frecuencia nos encontramos con que realizarlo no es posible. Esto podría deberse a que las "nuevas" secuencias presentan patrones muy divergentes (principalmente a nivel de secuencia) que no pueden ser analizados/comparados por métodos tradicionales, o a que frecuentemente no se cuenta con el templado adecuado para hacer un modelado de homología (*homology modelling*), debido a la falta de una estructura tridimensional resuelta para poder comparar. Dentro de las secuencias que con frecuencia imponen ambas dificultades, tenemos a los genes con tasa de mutación alta que difieren significativamente a nivel de secuencia (identidad menor al 35% aún entre familias proteicas) y a las secuencias que son llamadas "huérfanas" (que carecen de homólogos en otros linajes). Ambos tipos de secuencias están presentes con una frecuencia alta en las bases de datos y presentan obstáculos notorios, en los cuales la detección de similitud se complica, puesto que, en muchas ocasiones, estas secuencias/proteínas no son incluidas en los estudios (principalmente por la incertidumbre que formulan) o porque simplemente no pueden ser correctamente identificadas y anotadas, lo que a su vez implica que frecuentemente quedan registradas en las bases de datos como: secuencias no-anotadas, putativas, teóricas, no clasificadas o desconocidas.

Dado que la comparación, inclusión y categorización sistemática de estas secuencias no-anotadas es fundamental para entender de forma global diversos aspectos biológicos -pues es de suponerse que podrían tener implicaciones importantes-; el problema necesita abordarse, analizarse y resolverse de la manera más rápida posible. Para tal propósito, nosotros utilizamos el marco teórico del método de análisis de los clusters hidrofóbicos (HCA); considerado como una herramienta diagnóstica eficaz para identificar similitud (homología) en secuencia y/o estructura aún entre proteínas altamente divergentes (<20% de identidad).

Como el nivel de experiencia requerido por el usuario para interpretar los diagramas de HCA (unidad de análisis del método de HCA) es el factor más limitante para la utilización y uso generalizado del mismo, en este trabajo empleamos el marco teórico del método de HCA en conjunción con el algoritmo HCA_Analyze_Multiple_Aligns (HCAmA) para hacer una evaluación numérica/estadística de la significancia en las comparaciones automatizadas de diversas secuencias de acuerdo a la identidad entre agrupamientos (*clusters*) hidrofóbicos que se encuentran en las proteínas en estudio. Gracias a que el algoritmo sólo requiere de un alineamiento de secuencias para trabajar, un universo muy grande de proteínas es analizable, así que enfocamos nuestros esfuerzos en un caso bien conocido por la dificultad que implica estudiar su extensa variabilidad: Las Lisozimas de Bacterias. Haciendo una identificación de los Patrones Distintivos de Hidrofobicidad (PDH) que se encuentran en las secuencias de éstas enzimas, pretendemos comprobar que las Lisozimas, a pesar de tener estructura primaria muy divergente, poseen características comunes (principalmente estructurales) que son de utilidad en su identificación, comparación y clasificación, pues cuando éstos patrones distintivos (PDH) se mapean sobre un dendograma estructural, indican agrupaciones entre diversas secuencias que favorecen el entendimiento de las relaciones de ancestría/descendencia entre éstas enzimas. La metodología que aquí describimos es eficiente, rápida y sencilla, y aun así, lo suficientemente poderosa como para permitirnos visualizar gráficamente las agrupaciones que se forman entre éstas proteínas en base a su estructura (característica mejor conservada durante la evolución), con lo que podemos resumir, de una forma práctica y concreta, una vasta cantidad de información, característica que se puede aprovechar para efectuar comparaciones estructurales a gran escala que resulten robustos, completos y significativos.

1.1 Abstract

With the explosive increase in the number of coding sequences that come from the various sequenced genomes, there has been a problem of assigning identity to this vast amount of information accumulated day by day. A practical way to facilitate and accelerate the categorization of this "new" information is the systematic comparison of the “newly” obtained sequences against already annotated ones. Although this comparative exercise involves only the automation of the process, we often find that doing this is not always possible. This could be because the "new" sequences are very divergent (mainly at the sequence level) and cannot be analyzed/compared employing traditional methods, another reason it´s because, usually we don´t have adequate templates for homology modeling, mainly because of the lack of a resolved three-dimensional structure for multiples comparisons. Within the sequences which frequently involve both difficulties, we have genes with high mutation rates which differ significantly at sequence level (less than 35% identity even among protein families), and sequences called "orphans" (genes lacking homologues in other lineages). It´s reported that both types of sequences are present at high frequency in the databases and they represent noticeable obstacles for comparative analysis, in which the detection of similarity is complicated, since, in many cases, these sequences/proteins are not included in the studies (mainly because the uncertainty associated with the results) or because they simply cannot be properly identified and recorded, which in turn, implies that often are registered in the databases as non-annotated sequences, putative, theoretical, unclassified or unknown.

As the comparison, inclusion and systematic categorization of these non-annotated sequences is critical to understand globally diverse biological process -because it is assumed that they could have important implications-, the problem needs to be addressed, analyzed and resolved in the best way and as quickly as possible.

For this purpose, we use the theoretical framework of the method of Hydrophobic Cluster Analysis (HCA); regarded as an effective diagnostic tool for identifying similarity (homology) in sequence and/or structure, even among highly divergent proteins (<20% identity).

As the level of expertise required by the user to interpret HCA plots (the unit of analysis of the HCA method) is the most limiting factor for the widespread use of the method, in this paper we use the theoretical framework of the method of HCA HCA_Analyze_Multiple_Aligns in conjunction with the algorithm (HCAmA) to make an automated numerical/statistical evaluation of the significance in the comparison of sequences according to the identity of hydrophobic clusters that are found in the sequences of the studied proteins.

Because the algorithm only requires a sequence alignment to work, a very large universe of proteins can be analyzed, so, we focus our study efforts on a well known case for the difficulty of its extensive variability: The Bacterial Lysozymes.

Making an identification of the Distinctive Patterns of Hydrophobicity (DPH) in the sequences of these enzymes, we intend to verify that, despite possessing very divergent primary structure, DPH can be used as common characteristics indicators (mainly structural) that are useful in the identification, comparison and classification of these Bacterial Lysozymes. We also show that, when these distinctive patterns (DPH) are mapped onto a structural dendrogram, they can be used to indicate associations between a large group of sequences, which, in turn can promote understanding of the relationships of ancestry/descent and variability between these enzymes.

The methodology described here is efficient, quick and simple, and yet powerful enough to allow us to graphically visualize the groupings that are formed between these proteins based on their structure (best feature conserved during evolution). Such a vast amount of information is summarized in a practical and concrete way, a heavy advantage that can be exploited to perform large-scale structural comparisons that are robust, comprehensive and meaningful based solely in the study of sequence.

2. Introducción

2.1 Estructura de Proteínas

Las proteínas son polímeros compuestos por 20 aminoácidos diferentes unidos entre sí mediante enlaces peptídicos. En solución acuosa, bajo condiciones fisiológicas de presión y temperatura, las cadenas polipeptídicas de las proteínas suelen plegarse para formar -en la mayoría de los casos- estructuras globulares. Antes de adquirir su configuración tridimensional característica, las proteínas tienen diferentes niveles de organización, incrementado cada vez más la complejidad de las interacciones tridimensionales entre sus aminoácidos componentes. Siendo así, decimos que las proteínas tienen cuatro niveles de organización:

Primaria: Es la secuencia linear de residuos de aminoácidos que conforman la proteína.

Secundaria: Se forman segmentos locales plegados que se estabilizan por puentes de hidrógeno e interacciones electroestáticas débiles que siguen patrones de torsión definidos para formar los enlaces covalentes (enlace peptídico) entre los extremos amino y carboxilo de los aminoácidos que conforman la secuencia de la proteína. Las estructuras conocidas como alfa-hélices y hojas beta plegadas son características del segundo nivel de organización proteico.

Terciaria: Es el “acomodo” espacial de las estructuras secundarias para formar un plegamiento global y así dar lugar a una estructura tridimensional que define a las proteínas principalmente globulares.

Cuaternaria: Ensamblaje de subunidades, ya sean idénticas o distintas que conforman a una sola proteína.

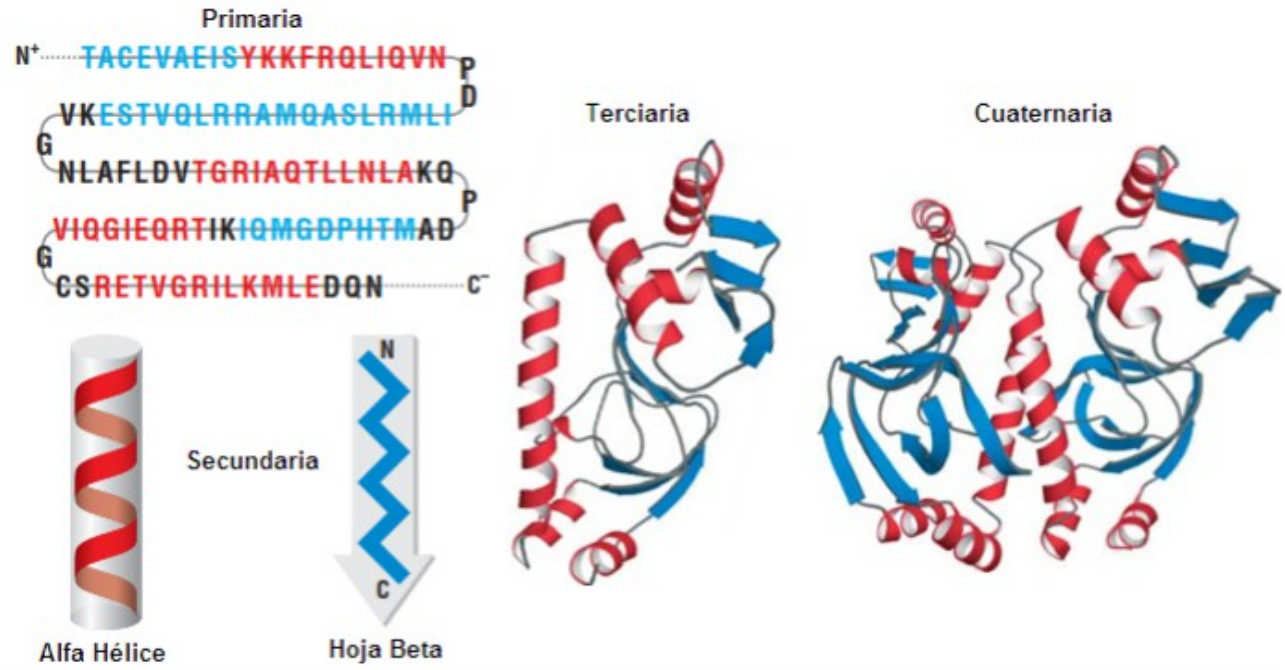

Figura 1. Los cuatro niveles de organización estructural en las proteínas. Modificaciones hechas a la imagen de la Referencia 38, página 3.

Para que un polipéptido funcione como una proteína, usualmente debe ser capaz de formar una estructura terciaria estable bajo condiciones fisiológicas (38). Al mismo tiempo, otra característica importante para la función de las proteínas es que no deben ser demasiado rígidas. Como discutiremos más adelante, se sabe que el universo de estructuras tridimensionales que las proteínas pueden adoptar es grande, aunque finito (38, 73). Si éste número limitado de conformaciones tridimensionales es consecuencia de factores puramente físico-químicas o de fuertes presiones de selección durante la evolución aún no ha podido definirse (38, 39, 43).

2.2 La importancia de los aminoácidos en la estructura y función de las proteínas

Las cadenas laterales de los aminoácidos suelen participar en interacciones entre ellas mismas, con moléculas de agua y con aquellas otras que se encuentran en el medio circundante. Las diferencias en las capacidades de interacción entre cada uno de los distintos aminoácidos influencian profundamente la estabilidad de las proteínas y su función. Aunque muchas características se pueden emplear para clasificar a los componentes estructurales de las proteínas, el carácter polar o no polar de los aminoácidos que poseen es notablemente el factor más determinante en las capacidades de cada uno para formar estructuras tridimensionales (38, 73).

Los aminoácidos hidrofóbicos, Alanina, Valina, Fenilalanina, Leucina e Isoleucina participan únicamente en interacciones de van der Waals. Son no-polares y tienen una marcada tendencia a evitar el contacto con el agua, favoreciendo el empaquetamiento entre sí. Estas agrupaciones entre aminoácidos hidrofóbicos son la base del denominado efecto hidrofóbico, que es uno de los aspectos fisicoquímicos de mayor peso en la formación del núcleo o "core" proteico de la mayoría de las proteínas globulares (38, 50). Dentro de los aminoácidos no-polares, la Alanina y la Leucina favorecen las estructuras tipo hélice, mientras que la Prolina, que tiene carácter anfipático, suele limitar la formación de hélices, ya que su estructura en forma de anillo produce impedimentos estéricos y limita su capacidad de formar enlaces por puente de hidrógeno con otros residuos de aminoácidos. La fenilalanina, un aminoácido muy hidrofóbico, al contrario, gracias a su estructura en forma de anillo (aromático) aún posee la capacidad de participar en interacciones polares débiles (38).

Los aminoácidos hidrofílicos: Arginina, Acido Aspártico, Glutámico, Serina, Treonina, Cisteína, Asparagina, Glutamina e Histidina son, en su mayoría, de carácter polar, aunque algunos de ellos se pueden clasificar como "neutros" o miméticos (que pueden ocupar posiciones normalmente reservadas para aminoácidos hidrofóbicos) en condiciones variables de pH. Estos aminoácidos son capaces de formar enlaces tipo puente de hidrógeno entre ellos mismos, con moléculas de agua, con moléculas orgánicas y con iones.

Como lo mencionamos anteriormente, los aminoácidos hidrofílicos poseen capacidades diferentes para formar enlaces, ya sea donando o aceptando electrones, esto dependiendo, principalmente, de las condiciones variables de pH. La Histidina, es posiblemente uno de los aminoácidos más versátiles en éste sentido, pues su estructura conformada por grupos N-H que pueden compensar la pérdida de protones le dan la posibilidad de pasar de ser ligeramente negativo, neutro o positivo, lo que biológicamente se traduce en una versatilidad aprovechada en los sitios activos de las proteínas.

Los aminoácidos considerados como anfipáticos, Lisina, Tirosina, Metionina y Triptófano también pueden modificar su PKa en condiciones variables de pH; sin embargo, la peculiaridad de sus cadenas les hace predominantemente no-polares en condiciones fisiológicas. La Metionina, por ejemplo, también es considerada como ligeramente hidrofóbica, dado que su cadena lateral es muy poco reactiva en comparación con el resto de los aminoácidos anfipáticos. Notablemente, el sulfuro de la metionina le permite coordinarse perfectamente con iones metálicos (38).

Los aminoácidos son componentes cruciales de las células no sólo por su versatilidad, sino porque polimerizan de forma muy sencilla y sin el requerimiento neto de energía (38). Gracias a ésta característica, las cadenas polipeptídicas se forman por la unión covalente de aminoácidos mediante enlaces tipo amida, frecuentemente denominados enlaces peptídicos. Los enlaces peptídicos se dan entre el extremo amino de un aminoácido y el extremo carboxilo de otro con la pérdida de una molécula de agua durante el proceso (38). Aunque es sabido que la hidrólisis de los enlaces peptídicos es una reacción exergónica, los enlaces peptídicos son muy estables en solución, esto es así porque la energía de activación para romper un enlace peptídico en solución acuosa es muy alta, así que en condiciones naturales o intracelulares esto ocurre a velocidades muy lentas, por lo que dan la impresión de ser muy estables. In vivo, este tipo de reacciones se facilitan gracias a las enzimas.

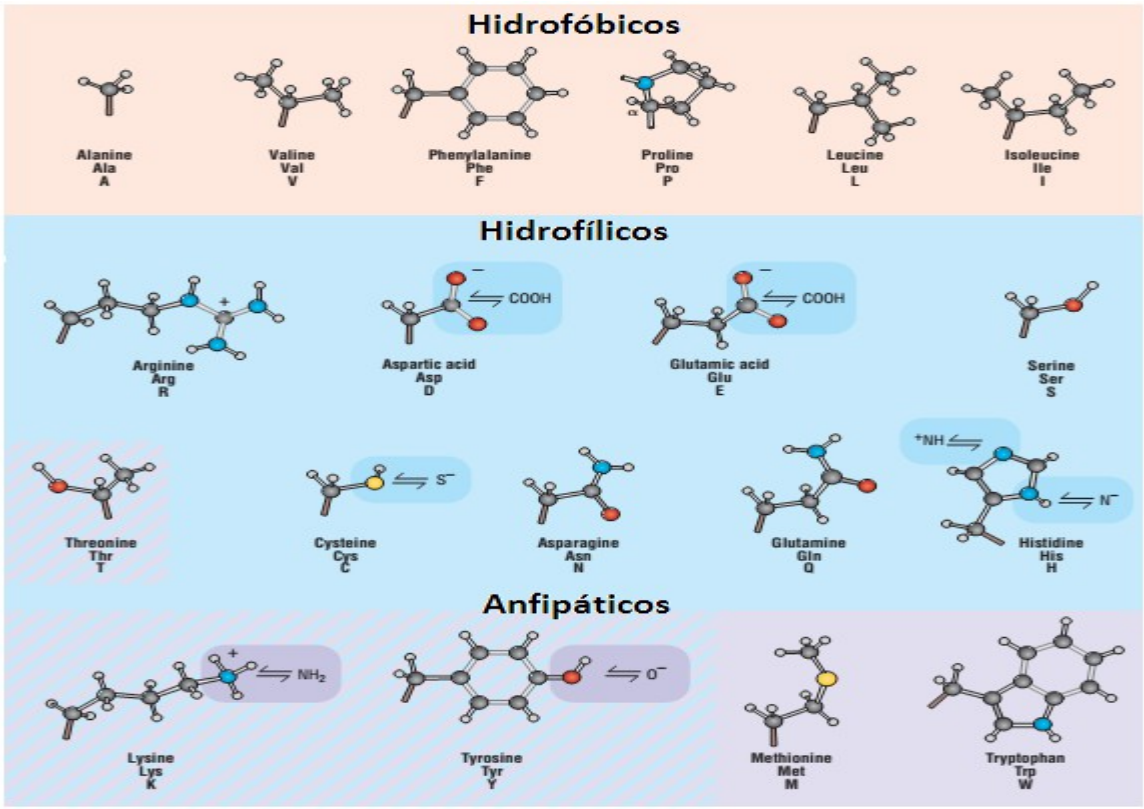

Figura 2. Los veinte aminoácidos y su clasificación en polares, no-polares y anfipáticos. Modificaciones hechas a la imagen de la Referencia 38, página 5.

Las proteínas se pueden formar por combinaciones de, al menos, 20 aminoácidos, aunque, ocasionalmente, pueden encontrarse aminoácidos modificados post-traduccionalmente (38, 73). Estas modificaciones suelen conferirles propiedades únicas que tienen diferentes propósitos. Algunos ejemplos de estas modificaciones son las carbo-amilaciones de la Lisina, que le permite

unir o utilizar iones metálicos, lo que a su vez funciona como un "*switch*", prendiendo o apagando una enzima. Otro ejemplo es la desaminación de la Asparagina, que puede alterar la estabilidad proteica e inducir cambios conformacionales bruscos en la misma. Este tipo de modificaciones a los aminoácidos son relativamente comunes en las proteínas de las plantas; sin embargo, en mamíferos sólo ocurre en algunas hormonas (6, 8, 38, 39, 40, 43).

2.3 Consideraciones importantes en el análisis de la estructura tridimensional

Una de las áreas de investigación más activas de la biología es la predicción de las conformaciones tridimensionales de las biomoléculas. Esta labor, que en la actualidad se enfrenta desde un ángulo multidisciplinario (combinando distintas ciencias, como la Matemática, la Física y la Química) ha resultado complicada, pues predecir con un buen nivel de confianza la estructura tridimensional de una proteína partiendo sólo de la secuencia de la misma -a pesar de que, en teoría, ésta contiene la información necesaria para determinar la conformación tridimensional que adoptará- aún está lejos de lograrse (73).

Considerando la velocidad a la que se secuencian genomas completos, hacer genómica/proteómica comparada, podría, de hecho, ser la única forma de estudiar el genoma de los organismos que secuenciamos. Esta labor de comparación se realiza, primordialmente, a dos niveles: secuencia y/o estructura. En el primer caso, se incluyen métodos de "muestreo o enhebramento" (*threading*) y modelado comparativo de estructura, también conocidos como modelado de homología (*Homology Modelling*); el cual se define como un método computacional para modelar la estructura de la proteína basada en la similitud de secuencia/estructura con una o más proteínas de estructura conocida. Los casos en los que se cuenta con una estructura tridimensional resuelta de la proteína de interés (así como de proteínas similares) generalmente se utilizan como modelos para poner a prueba el poder predictivo de los algoritmos empleados en tales propósitos; lo que nos lleva a un segundo caso, en los cuáles se recurre a métodos que hacen una predicción del plegamiento/estructura de las proteínas utilizando sólo la información de la secuencia. Estos métodos no se guían con base a estructuras resueltas producidas por secuencias idénticas a la secuencia problema. Las simulaciones del plegamiento de las proteínas toman en cuenta diversas características de las secuencias que les producen; por ejemplo, toman en consideración la estructura secundaria, la propensión de los aminoácidos a formar estructuras determinadas, e incluso, la naturaleza misma de las proteínas; en cuyo caso, se espera que una

proteína extracelular posea una proporción más elevada de lo normal del aminoácido Cisteína (38).

La predicción de estructuras tridimensionales a partir de la comparación directa con estructuras similares o la predicción de la estructura o plegamiento a partir de la secuencia, ha adquirido una enorme relevancia en la actualidad, pues es la forma más sencilla, rápida y barata de explorar y comparar la información que se acumula en grandes cantidades en las bases de datos (38, 41, 52). Los métodos para comparar, hacer búsquedas en bases de datos o predecir estructuras son variados y hacen uso de distintas representaciones, algoritmos puntuadores, de optimización, y de categorización. El común denominador en los dos primeros casos es que tales algoritmos explotan las identidades que se esperarían a nivel atómico entre estructuras similares; mientras que al mismo tiempo, cálculos de "puntuación" (*scores*) basados en diversas formulaciones matemáticas, son los que evalúan la exactitud de las comparaciones, mismas que se utilizan como parámetros de discriminación para evaluar alineamientos estructurales o hacer súper-imposiciones estadísticamente significativas y así lograr diferenciar entre posibles falsos positivos. Tradicionalmente, los algoritmos discriminadores se basaban en un conocimiento empírico y/o experimental, y aunque se ha dado el caso en que algunos producen resultados paradójicos, por su enorme utilidad, son frecuentemente implementados en estos métodos (41).

A pesar de los avances, el alcance de los algoritmos que tratan de predecir la estructura tridimensional de una molécula con frecuencia se ven limitados; pues en muchas ocasiones no se tiene la estructura tridimensional resuelta de la molécula, o al menos, una representante de una familia proteica contra los cuales se pueda evaluar una hipótesis de estructura. En estos casos, las inferencias se hacen partiendo sólo de la secuencia de nucleótidos/aminoácidos de la proteína misma, esto es, cuando no hay secuencias similares con estructuras resueltas para comparar, sólo los modelos *de novo* o *ab initio* se pueden aplicar.

La determinación de la estructura tridimensional de las biomoléculas, partiendo únicamente de la secuencia, asume varios escenarios considerados como plausibles pero que no han sido posibles de corroborar experimentalmente (38). Por ejemplo, se asume que la forma plegada (de las proteínas) equivale a la conformación en equilibrio y de menor energía, que a su vez es la cantidad mínima de energía que requiere la proteína para no desnaturalizarse. Se asume, además, que los cálculos que se llevan a cabo para medir la contribución de las fuerzas electroestáticas así como para predecir las interacciones moleculares entre los átomos que participan, son lo

suficientemente precisas como para “acercarse a la realidad”, pero hasta el momento, las certezas de tales predicciones no ha podido ser comprobada (38). Tampoco se ha podido evaluar cuándo una secuencia producirá una proteína oligomérica; siendo que la predicción de estructura, en la actualidad, se basa solamente en abordar a las proteínas monoméricas (38). Todavía más intrigante, es el hecho de que no se sabe cómo se da el plegamiento de las proteínas asociadas a membrana. Este tipo de proteínas muestran conformaciones estructurales limitadas en comparación con las globulares (38). Dado que éstas no se encuentran en medios acuosos, se piensa que las interacciones intramoleculares tempranas (entre los aminoácidos propios de la proteína y los no-polares que se encuentran en las regiones internas de las bicapas lipídicas, por ejemplo) son las responsables de guiar el plegamiento en cuanto se forma la proteína; sin embargo, los detalles de éste fenómeno son inciertos (38, 39, 52).

Existen dos aproximaciones comunes para hacer modelados de homología (*Homology Modelling*). En primera está el ya mencionado muestreo basado en perfiles (*threading*), que consiste en forzar una secuencia de estructura desconocida a adoptar todos los tipos de plegamiento conocidos. La eficacia de la aproximación se mide en función de qué tan bien se ajusten los distintos plegamientos a las características dictadas por las secuencias La búsqueda se hacen en un universo de posibles conformaciones tridimensionales en las que la energía libre es particularmente baja para una proteína con una correspondiente secuencia de aminoácidos. Naturalmente, es factible obtener más de un posible resultado, así que el algoritmo de igual forma evalúa la probabilidad de que una secuencia adopte determinado plegamiento mediante el parámetro Z, que considera la media de ajuste de secuencias al azar que de igual forma podrían producir una estructura similar a las obtenidas. Si un número “n” de estructuras favorecen condiciones que se cumplen en proteínas conocidas (por ejemplo, que los aminoácidos hidrofóbicos se encuentren en el interior y los hidrofílicos en el exterior) aumenta el valor del parámetro Z, lo que se puede interpretar como que la secuencia es capaz de producir una proteína que adopte tal plegamiento, más no que sea su conformación nativa o de energía libre mínima (38). Como podemos ver, el cómputo de parámetros fácilmente puede volverse restrictivo, así que para acelerar las búsquedas del universo de posibles conformaciones tridimensionales, sólo se consideran de manera explícita a ciertos átomos de las cadenas proteicas, el resto son modelados matemáticamente o no son considerados explícitamente; de esta manera los átomos no considerados (que pueden ser parte de la estructura o pertenecer a moléculas presumiblemente importantes para el plegamiento de las proteínas, como el agua) pueden ser expresados mediante promedios (38, 42).

Dada la naturaleza de éstas búsquedas heurísticas (promediadas), podemos notar que hacer inferencias *de novo* en secuencias divergentes fácilmente podría arrojar resultados no concluyentes, inexactos o equivocados (38, 42). A pesar de las limitaciones, estos métodos se siguen usando y poseen cierta ventaja, pues se ha confirmado que cuando las secuencias tienen una identidad menor al 30%, los métodos comparativos de igual forma son muy propensos a errores (52, 41, 38).

Una segunda aproximación más prometedora es la de *Rosetta*, que fue un método desarrollado dentro de la competencia del CASP (*critical assessment of techniques for protein structure prediction*) cuyo propósito es predecir la estructura tridimensional de proteínas que han sido resueltas (se conoce su estructura de forma experimental) pero que no se han publicado (52, 38).
El principio de *Rosetta* es ajustar segmentos regulares de la proteína. Estos segmentos corresponden a estructuras secundarias que pueden ser inferidas fácilmente (y con un buen nivel de confianza) partiendo de la secuencia (38, 73). Mediante comparaciones de estos segmentos con otros de secuencia parecida que forman estructuras secundarias similares, se guía poco a poco hacia un plegamiento tridimensional completo en el que una vez más la energía libre mínima es el parámetro de mayor peso para hacer aproximaciones hacia el plegamiento más probable (73). Aunque cada segmento de la proteína es evaluado individualmente, la búsqueda de posibles conformaciones se acelera notablemente, pues se hace tomando en cuenta un universo de soluciones limitadas (que se extrapola de la energía libre que presentan estructuras conocidas y resueltas experimentalmente) en el que un pequeño cambio en un segmento puede repercutir significativamente en los demás (52, 73). Así, el ajuste (orientación en el espacio) de segmentos cortos aunque completos, de estructura secundaria ya conocida, pero limitado a valores determinados experimentalmente, es lo que hace a Rosetta un método rápido y confiable. Lamentablemente, dado que Rosetta utiliza la identidad de secuencias para buscar y predecir el comportamiento de estructuras secundarias, la similitud de la secuencia con otras ya reportadas en las bases de datos puede convertirse en un factor determinante para poder inferir la estructura tridimensional más plausible o correcta (38, 52, 73). Aun con los resultados positivos en la predicción de estructuras *de novo*; por su confiabilidad, suele abogarse a favor del modelado de estructuras homólogas (*homology modelling*), aunque, como mencionamos, la falta de secuencias representantes de diversos dominios proteicos son el impedimento más significativo para hacerlo.
Experimentos en los que se analizan y comparan las estructuras de las más diversas moléculas han arrojado una gran cantidad de evidencia que apoya la noción de que la estructura tridimensional

de una proteína suele conservarse mejor que la secuencia (6, 8, 38, 50, 66, 73); y, aunque las razones para este hecho son variadas, es factible hipotetizar que la naturaleza conservativa del código genético (que amortigua los cambios) sea de primordial importancia (38, 73). Uno pensaría que la conservación de la estructura, que es importante en la función, implica un origen común, pero, en ocasiones, una estructura similar no necesariamente implica homología, y, de forma más dramática, secuencias similares no siempre producen plegamientos equivalentes, así como secuencias similares en ocasiones dan lugar a plegamientos completamente distintos (38, 73).

A pesar de esto, los resultados obtenidos en la predicción de estructura tridimensional ya alcanzan una precisión del 70% (con proteínas globulares de un sólo dominio) [ver Referncia 38 para más detalles], en los que el modelo predicho coincide con la estructura tridimensional "real", y aunque esto resulta variable (para algunas estructuras la precisión y/o resolución es menor) diversas limitaciones metodológicas, interacciones globales no consideradas (presión, pH, temperatura, solventes, etc. dentro de la célula), fuertes restricciones al estudio de proteínas multidominio, y todo lo anterior aunado a que, ciertamente, pueden existir algunos otros parámetros que tal vez aún desconocemos, son los factores que, con mayor frecuencia, limitan la precisión de las predicciones de estructura y/o función proteica (38, 41, 52, 73).

Desde hace algún tiempo y a pesar de la problemática, los extensos estudios en el tema ya han logrado establecer con un buen nivel de confianza que, debido a que la estructura tridimensional que tienden a adquirir aún las más diversas proteínas está restringida por propiedades físico-químicas (siendo el ángulo de los enlaces atómicos a los carbonos centrales (alfa) en relación con el enlace peptídico uno de los más importantes), el universo posible de plegamientos proteicos está restringido a sólo unas cuantas conformaciones tridimensionales [es posible que sólo en el orden de los millares, pues los datos con los que se contaba en 2004, indicaban que el universo de plegamientos posibles distintos era de aprox. 1,000 a 10,000. (Ver Ref. 57 para más detalles) (38, 64)].

Según lo anterior, es factible pensar que incluso las proteínas que a nivel de secuencia sean altamente divergentes, posean la capacidad de adoptar estructuras tridimensionales muy semejantes a otras ya conocidas y posiblemente anotadas, lo que facilitaría considerablemente la labor de identificación y clasificación de las mismas (5, 38, 43, 50). Esta información, que desde su manera más primordial y sencilla puede ser pre-evaluada en base a los diagramas de Ramachandran para los valores de rotación/torsión en los ángulos *phi* (φ) y *psi* (ψ) de los enlaces

entre los átomos C-Alfa, N-H (el enlace peptídico) y su relación con el grupo R de los aminoácidos presentes en las proteínas (38), es una de las evidencias teóricas más fuertes que da sustento a otra de las ideas fundamentales en el estudio comparativo de la estructura de las proteínas, y esto es el descubrimiento de que una gran variedad de proteínas -aun aquellas de las más diversas funciones-, están agrupadas dentro de las mismas familias homo/polifuncionales, compartiendo al menos un elemento de su estructura (p. ej. una agrupación distintiva de alfa-hélices e inclusive un dominio estructural, que a su vez suelen ser producto de características distintivas pero compartidas), lo que incrementa la confianza para analizar la información nueva con base a comparaciones con lo ya conocido (6, 8, 38, 43, 73).

Aunque el muestreo ("*screnning*") de las bases de datos y los métodos para hacerlo (heurísticos, en principio) es un área de investigación y refinamiento muy activa, aquí no nos adentraremos más, pues nuestro interés primario es explorar a fondo una de las opciones que consideramos más promisorias para el análisis de la estructura de proteínas (independientemente de que exista la estructura resuelta de nuestras enzimas de estudio); es decir, la disposición espacial de los agrupamientos (*cluters*) de aminoácidos hidrofóbicos VILFMYW, partiendo sólo de la información más básica y sencilla con la que normalmente se cuenta: la secuencias (4, 5, 38, 50).

2.4 Los aminoácidos hidrofóbicos como "caracteres estructurales" en la comparación de proteínas

Desde que se comenzaron a secuenciar proteínas, surgió un interés primordial en conocer y predecir la estructura de las mismas partiendo sólo de la secuencia de aminoácidos. Desde trabajos que datan del final de los años 50, se sugería que las interacciones hidrofóbicas (que primordialmente sólo se pueden dar entre aminoácidos hidrofóbicos) eran importantes para el establecimiento del plegamiento proteico (4, 58). Conforme se realizaron experimentos y la información incrementó, se obtuvieron suficientes datos empíricos que apoyaban la idea de que habría dos tendencias inversas pero no independientes que se podían ver reflejadas en la estructura tridimensional de las moléculas (58, 59, 60, 66). Nos referimos a la predisposición de los aminoácidos hidrofóbicos a concentrarse y aislarse de las zonas de contacto con moléculas polares (como el agua); mientras que, por el contrario, los aminoácidos hidrofílicos, suelen ocupar sitios de la periferia molecular en donde se maximizan el número de interacciones (58, 59). Tiempo después, fue evidente que existían ligeras variaciones a tal fenómeno, por lo que ahora sabemos

que es posible encontrar aminoácidos hidrofóbicos en la periferia o hidrofílicos en el núcleo proteico (8, 58, 60).

La tendencia a ocupar determinada posición en la estructura tridimensional de una proteína no sólo está dictada por el carácter polar de un aminoácido; sino que otras características físico-químicas, como los impedimentos estéricos, la reactividad (que tan fácil reaccionan) e incluso su papel único en determinadas reacciones químicas (como el grupo tiol de la Cisteína, por ejemplo) son factores que también poseen una importancia relativa en restringir o favorecer su presencia en posiciones en las que normalmente no se esperaría observarlos. Aun considerando estas excepciones, ciertamente se pudo corroborar que la polaridad de los aminoácidos era un buen parámetro discriminatorio y matemáticamente correlacionado con las regiones "ocultas o internas" y que ayudan a identificarlos de las que son expuestas en las proteínas (58, 59). Acentuando todavía más el alcance predictivo y analítico que la polaridad de los aminoácidos podía tener en el estudio de la secuencia y/o estructura proteica, se crearon las escalas de hidrofobicidad (58, 59, 60); las cuales tenían el propósito de facilitar la identificación de regiones importantes para la adopción de la estructura tridimensional final de las proteínas en estudio.

Desde aquí, es importante resaltar que las escalas de hidrofobicidad no son una característica intrínseca al método de HCA, en primera instancia porque no todos los 20 aminoácidos son considerados y no todos poseen el mismo peso; sin embargo, es notorio que el alfabeto del método (los aminoácidos considerados como fundamentales dentro del marco teórico del mismo. Ver más adelante para más detalles) asigna "pesos" diferenciales para evaluar de una forma más realista su contribución en la formación de estructuras secundarias, lo que es concordante con la lógica detrás de las escalas de hidrofobicidad usadas en otras metodologías (38, 58, 64).

A finales de los años 70s, trabajos comparativos como los de Margaret Dayhoff, en los que se estudiaban una gran cantidad de secuencias proteicas para determinar la variabilidad en ellas -laborioso trabajo empírico que eventualmente culminaría en las matrices de sustitución de PAM-, confirmaban indirectamente que las características distintivas de las secuencias que se traducían a proteínas eran más similares dependiendo del nivel de identidad entre las mismas (64). Esto es, proteínas que pertenecían a la misma familia homofuncional o de especies filogenéticamente emparentadas eran producidas por secuencias de características similares, algo que hoy en día se intenta explotar en comparaciones a muchos niveles, en lo que suele denominarse como análisis

de firmas genómicas (38, 73). Margaret Dayhoff y colaboradores, clasificando las secuencias en familias y súper-familias, aplicando métodos estadísticos a la comparación minuciosa de segmentos estructurales -bajo la hipótesis de que secuencias con menor identidad sólo coincidirían en segmentos pequeños- intentó clasificar a las proteínas bajo una lógica muy similar a lo que hoy hacemos cuando buscamos dominios o motivos (38). W. F. Doolittle (58) y Dayhoff (64) pudieron notar que, aún las secuencias con identidad menor al 20% podrían tener un origen común, por lo que encontrar patrones, regiones, e incluso aminoácidos conservados (Dayhoff proponía que la Fenilalanina, la Tirosina y la Cisteína, tres aminoácidos importantes dentro del marco teórico del HCA, eran buenos candidatos para analizar aminoácidos poco variables, pues no sólo sufren pocas sustituciones sino que, además, suelen hacerlo entre sí) era un paso fundamental para hacer comparaciones significativas entre secuencias divergentes o con poca similitud (64).

En la misma época, se propone que la comprensión de las propiedades de los aminoácidos hidrofóbicos (que incluyen su frecuencia y distribución) eran importantes para entender la estructura proteica (58), por lo que, apoyados en la gran cantidad de evidencia que así lo indicaba, se consolidó el estudio de los aminoácidos hidrofóbicos como "caracter estructural" para el análisis a diversos niveles de las proteínas y sus estructuras (ejemplificado en el desarrollo de técnicas comparativas basadas en los perfiles de hidrofobicidad de secuencias completas).

2.5 El Método del Análisis de los Agrupamientos (*Clusters*) Hidrofóbicos (HCA)

Así, gracias a que aun las más diversas proteínas comparten elementos a nivel de secuencia y estructura que podrían facilitar su clasificación, sumado a la confirmación de que las proteínas globulares se requieren de en un ambiente acuoso donde los aminoácidos hidrofóbicos juegan un papel fundamental para mantener el núcleo o "*core*" de la proteína y que, por ende, este tipo de aminoácidos pueden ser de gran utilidad en el estudio de las proteínas; el desarrollo de metodologías que aprovecharan tales cualidades fue inminente. Entre ellas, en 1987, se introduce el método del análisis de los agrupamientos (*clusters*) hidrofóbicos (HCA) por Christine Gaboriaud *et al* (4); el cual está basado en el análisis comparativo de las regiones hidrófobas constituidas por el agrupamiento de aminoácidos hidrofóbicos (*clusters*). Estas regiones son muy importantes e informativas, pues resultan determinantes en la predisposición para formar estructuras secundarias regulares por parte de cualquier proteína globular (4, 5, 6, 8, 38, 44, 50, 72).

La asociación de, al menos dos aminoácidos hidrofóbicos, se define con el nombre de

agrupamiento o *cluster*; por lo tanto, un agrupamiento o *cluster* es un conjunto de aminoácidos hidrofóbicos ininterrumpidos por aminoácidos polares (como G,D,N,S o Prolinas), que se encuentran muy próximos entre sí formando agrupaciones distintivas. Los clusters de aminoácidos hidrofóbicos se conectan, encierran y aíslan con el propósito de resaltarlos de forma evidente en la representación helicoidal de los diagramas de HCA.

Utilizar la naturaleza polar y la posición de los agrupamientos (*clusters*) de aminoácidos como caracteres "informativos" para analizar la estructura de las proteínas resulta altamente informativo, pues se ha demostrado que los aminoácidos hidrofóbicos V, I, L, F, M, Y, W, poseen un rol determinante formando el núcleo o core y participando activamente en la estructura secundaria de algunas proteínas (como las Alfa-hélices o las hebras Beta-plegadas) adoptando conformaciones estructurales tridimensionales conservadas durante el plegamiento proteico, lo que podría estar en relación con la función de la proteína, como las interacciones con ligandos y/o sustratos y/o con la presión de selección a la que los genes que les producen están sujetos (4, 5, 6, 8, 38). Se especula que por sus características de conservación, en las que este tipo de aminoácidos suelen experimentar únicamente sustituciones sinónimas o silenciosas (6, 8, 38, 54); para cualquier secuencia proteica, la prevalencia de aminoácidos hidrofóbicos representa un patrón bien conservado y correlacionado con la función biológica de las mismas; por tanto, su presencia puede considerarse como un indicador de similitud por origen común y no sólo un evento de coincidencia azarosa de poca significancia para éste tipo de estudios. Como veremos más adelante, el trabajo, desarrollo y análisis hecho en el área de la biología estructural ha permitido no sólo corroborar éste tipo de suposiciones, sino que, además, la búsqueda de patrones estructurales y su asociación con aspectos biológicos de las proteínas, que es una de las áreas de investigación más promisorias de la biología contemporánea (38, 50).

El estudio de los aminoácidos hidrofóbicos, que como mencionamos, incluye su hidrofobicidad relativa, su propensión, frecuencia y distribución, se ha convertido en un aspecto muy importante en los estudios genómicos y estructurales. Resulta interesante cómo, en la actualidad, se siguen sumando metodologías cuyo sustento teórico explota el uso como "marcador estructural" de los 8 aminoácidos más hidrofóbicos. Por ejemplo, en el trabajo de Faure y Callebaut del 2013, haciendo uso de una variante a los estudios de escalas de hidrofobicidad y empleando los diagramas del método de HCA como herramienta de análisis estructural, se delimitan posibles dominios y regiones desordenadas dentro de una secuencia proteica para hacer una caracterización de lo que

ellos denominan el “foldoma” de las proteínas (65). Además, el trabajo de Baussand *et al*. quienes utilizan el método de HCA como un control positivo para evaluar la precisión de los resultados en un estudio que intenta optimizar el alineamiento de secuencias divergentes (50).

2.6 Ventajas y desventajas del método de HCA

Existen una gran cantidad de sustentos metodológicos para hacer comparaciones a distintos niveles de dos o más secuencias cualesquiera. Todos ellos tienen ventajas y debilidades, por lo que hacer una descripción (aún general) de los mismos está más allá del objetivo de este escrito. Hasta ahora, podemos notar que el método de HCA es práctico para enfrentarnos al problema de identificar homología entre biomoléculas para las cuales sólo se conoce su secuencia o en las que el porcentaje de identidad (comparando contra las que se encuentran en las bases de datos) es bajo. Parte importante de éste trabajo es resaltar las bondades y ventajas del método, así como de señalar sus debilidades potenciales. En las siguientes líneas, quisimos favorecer un entendimiento preciso de las razones por las cuales el método de HCA puede ser muy práctico para analizar secuencia/estructura de proteínas.

1) Las secuencias de las proteínas suelen ser tolerantes a las sustituciones llamadas “*Indels*” (del inglés *insertions-deletions*) y demás cambios que puedan ocurrir. Estos pueden dificultar significativamente el trabajo de comparación, algo que con frecuencia se refleja en la dificultad para alinear las secuencias (41, 8).

Hay evidencia de que la estructura tridimensional de las proteínas suele conservarse mucho mejor que su secuencia, pues la organización del código genético favorece los cambios conservativos (38, 42, 50, 73). Por ejemplo, los cambios en tercera posición (del codón), usualmente no cambian el aminoácido; pero cuando éstos ocurren en segunda posición, con frecuencia cambian el carácter polar del aminoácido; sin embargo, algunos aminoácidos polares y no-polares pueden mimetizar las características fisicoquímicas entre ellos bajo condiciones especiales (de pH, principalmente, como la Lisina y la Histidina), lo que minimiza el efecto producido por los cambios a nivel estructural. Si por ejemplo, ocurre un cambio de un aminoácido polar (como la Serina) por uno no-polar (como la Valina), es posible que bajo condiciones variables de pH, la Serina puede fungir como aminoácido hidrofóbico y viceversa. Notamos entonces, que si el efecto se amortigua a nivel estructural, es muy factible que la función se conserve. Además, se ha propuesto que los *Indels* ocurren predominantemente en sitios que conforman regiones de interconexión entre dominos,

también llamados “asas” o "*loops*"; así como también en la periferia de las proteínas, posiciones que con frecuencia son muy variables y abundantes en aminoácidos hidrofílicos (4, 6, 8, 38). Por éstas razones, los métodos que trabajen a nivel estructural resultan especialmente útiles para realizar estudios comparativos (4).

2) Mediante los diagramas de HCA es sencillo hacer un “alineamiento estructural” de forma que se maximice la identidad estructural entre las secuencias/diagramas analizados, y con ello facilitar la identificación de regiones conservadas, independientemente de que podrían o no encontrarse en las mismas posiciones a nivel de secuencia (4, 6, 8). Esto es muy útil en el análisis de secuencias divergentes, pues normalmente, éstas secuencias se detectan y comparan sólo en pequeñas regiones (que corresponden a zonas o aminoácidos conservados) sin tomar en cuenta las regiones que son discordantes. Con los diagramas de HCA se pueden analizar aspectos como la orientación relativa de un asa o de una hoja-beta (antiparalela o paralela), que son cuestiones que normalmente suelen resolverse sólo si se tiene la estructura tridimensional a una buena resolución (menor a 2.5 nanómetro) de las proteínas en estudio.

3) Otra de las ventajas y que proviene del énfasis que se hace con el método de HCA en el análisis de la estructura secundaria, es que los residuos invariantes e hipervariables pueden ser identificados y examinarse a profundidad. Por ejemplo, es sencillo identificar los aminoácidos que participan en la actividad catalítica y que naturalmente suelen ser invariantes, aunque aquí nos conciernen más aquellos aminoácidos que tienen papeles preponderantes en la estructura tridimensional de las proteínas y que pueden ser analizados para determinar propiedades interesantes de éstas. Aplicando metodologías como el análisis de HCA se pueden obtener resultados importantes que antes sólo eran posibles mediante la aplicación de procedimientos que son más laboriosos y costosos, como los experimentos de mutagénesis dirigida (4, 6).

4) La combinación de metodologías suele ser una estrategia muy efectiva. Por ejemplo, hacer uso del método de HCA con métodos que alinean y/o predicen estructuras secundarias puede aumentar la capacidad del investigador para determinar las propiedades estructurales de alguna biomolécula. Más adelante en este trabajo, se retomará este punto, con el que explicaremos como utilizamos el método de HCA de una forma automática para poder hacer comparaciones a nivel estructural entre múltiples secuencias.

5) Hay proteínas que presentan retos notables para ser cristalizadas y muy posiblemente sólo se puedan analizar (de forma rápida, práctica y sin una gran inversión económica) con métodos como el de HCA. Como ejemplo, tenemos a la proteína que produce el gen *GvpA*, que es un elemento estructural de las vesículas de gas de algunas Arqueas. Esta proteína es considerada como una de las más hidrofóbicas que se han descubierto (51). Al ser muy hidrofóbica, métodos de purificación (que usualmente involucran solventes inorgánicos) y cristalización (algunas proteínas se "fijan" hidratándolas, pues las interacciones con las moléculas de agua les facilita adoptar su estructura tridimensional nativa [38, 77]) no funcionan con éste tipo de proteínas, así que los métodos bioinformáticos -al menos por el momento-, son la única manera de estudiar la estructura tridimensional de éste tipo de moléculas.

Consideramos que conforme se descubran más proteínas de membrana altamente hidrofóbicas, este problema tomará una gran importancia, pero es uno que se podrá resolver si se aplican metodologías que puedan hacer inferencias estructurales partiendo de sólo la secuencia, como es el caso del método de HCA.

6) Aunque en principio es más factible analizar proteínas en las que *a priori* se tenga conocimiento de que poseen aminoácidos hidrofóbicos importantes para el mantenimiento de su estructura o función, proteínas pequeñas -mayoritariamente polares-, en las que los puentes disulfuro son los encargados de mantener la identidad estructural de las proteínas, no son realmente factibles de analizar con éste método (4, 6, 38). A pesar de esto, hay que tener en cuenta que los diagramas de HCA muestran una distribución espacial relativa de todos los aminoácidos, incluyendo a los polares, por lo que el método de HCA podría ser una herramienta complementaria a métodos más idóneos para tales análisis.

7) El método de HCA funciona bien en los análisis de proteínas de diversos tamaños; sin embargo, esta cualidad también es una desventaja, pues se ha visto que el análisis de los diagramas de HCA de segmentos o de proteínas pequeñas pueden ser difíciles de analizar, arrojar resultados inconclusos o peor aún, producir falsos positivos (6). Esta desventaja puede superarse comparando sólo los segmentos más conservados entre familias proteicas o haciendo uso de una metodología que nos indique cuando las similitudes entre las secuencias se podrían deber solamente al azar.

8) Sólo hasta muy recientemente, los algoritmos de Pedro Silva (2009) han dotado al método de HCA de nuevas virtudes. La más importante, es que se disminuye sustancialmente el componente subjetivo del análisis de los diagramas de HCA; ya que ahora, un algoritmo puntuador es el encargado de seleccionar las mejores secuencias candidatas para ser inspeccionadas exhaustivamente, al mismo tiempo que es posible diferenciar entre los falsos positivos que pueda existir al hacer comparaciones pareadas entre secuencias. Esta nueva capacidad del método facilita que una gran cantidad de secuencias sean evaluadas bajo un mismo marco teórico, eliminando así -al menos de los primeros pasos- una de las críticas más fuertes al método, que es la subjetividad al realizar comparaciones entre los diagramas de HCA.
Reduciendo sustancialmente la subjetividad de las comparaciones, el método se vuelve accesible a un mayor número de usuarios que no requieren de un amplio conocimiento en bioquímica de aminoácidos y/o estructura de proteínas para obtener información sustancial y confiable de los análisis mediante el método de HCA.

A partir de ahora, continuaremos con la descripción de una parte muy importante del mismo, que es además, la unidad de análisis: Los diagramas o *plots* de HCA.

2.7 Los Diagramas de HCA

Los diagramas de HCA (Figuras 3 y 6) son una representación en dos dimensiones de una secuencia proteica. Los diagramas de HCA se hacen con el programa HCA-Draw de Luc Canard (79). Esta representación esquemática de la secuencia de aminoácidos que conforma una proteína muestra los aminoácidos siguiendo el plano de una Alfa-hélice desenrollada, cada vuelta correspondiente a aproximadamente 3.6 residuos; por tanto, el primer aminoácido de la secuencia corresponde al primero en el extremo superior izquierdo del diagrama, continuando con la lectura descendente en el plano vertical hasta llegar al cuarto aminoácido; cuando esto suceda, se continúa la lectura en la segunda columna del diagrama, así hasta hacer la cobertura completa de la secuencia (Figuras 3 y 6). El por qué se leen los aminoácidos de cuatro en cuatro tiene que ver con el desenrrollamiento artificial de la representación en Alfa-hélice, pues esto conlleva que algunos aminoácidos que originalmente estaban en contacto (en la estructura 3D de la proteína), se separen y queden representados como si en realidad estuviesen alejados, por lo que los diagramas muestran una duplicación o repetición de la disposición de los aminoácidos en el plano vertical del cilindro distendido, característica que es distintiva de las representaciones gráficas

hechas con éste método. La duplicación de un segmento de la secuencia tiene el propósito de proveer el plano para las dos dimensiones, la cual permite mejorar la apreciación de los patrones que los agrupamientos hidrofóbicos presentan a lo largo de la secuencia, y dar un panorama general de las interacciones cercanas o distales durante el plegamiento de la proteína, que como mencionamos, es una manera de compensar la separación "artificial" que ocurre en los diagramas [Figura 3]. Considerando la duplicación de los aminoácidos y la extensión de los mismos en el plano vertical, los diagramas de HCA muestran aproximadamente 5.3 aminoácidos por vuelta de Alfa-hélice distendida. (5, 6, 8, 44).

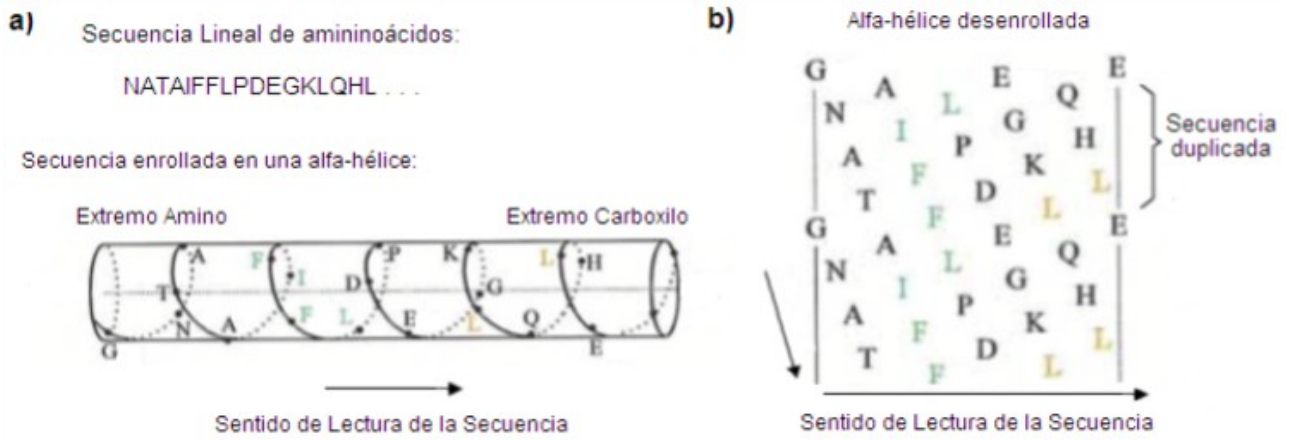

Figura 3. a) Aquí se muestra el principio fundamental de los diagramas de HCA. Una secuencia lineal se puede escribir siguiendo el plano de una alfa-hélice con 3.5 aminoácidos por vuelta (cilindro). Este se puede "desenrrollar" y duplicar en un plano vertical, con lo que se trata de representar en dos dimensiones, la ubicación espacial que cada aminoácido ocuparía en la estructura tridimensional. **b)** La segunda parte del esquema muestra en qué dirección se lee la secuencia (flechas) así como el segmento que se ha duplicado y no debería considerarse como parte original de la misma. Adaptado y Modificado de las Referencias 6 y 8.

La distancia que separa a los agrupamientos de aminoácidos hidrofóbicos, así como a la capacidad de discernir cuando un aminoácido pertenece a diferentes agrupamientos está definida por la distancia de conectividad [Figura 4]. Esta distancia corresponde al número de residuos polares y/o indiferentes al entorno (que pueden comportarse como hidrofóbicos si se encuentran dentro de un ambiente no-polar), que separan dos agrupamientos hidrofóbicos identificados como individuales dentro de las representaciones en alfa-hélice. La distancia de conectividad (que no es un concepto propio o único del método de HCA; sino que es una manera convencional de hablar de las posiciones relativas de los aminoácidos en las representaciones de alfa-hélices implementadas también en otros métodos), tiene dos aplicaciones importantes. La primera es que nos permite observar agrupamientos de aminoácidos hidrofóbicos individuales, cada uno con

características propias (y con implicaciones durante el plegamiento) que no se traslapan entre ellos, lo que favorece la interpretación e identificación de patrones en los diagramas de HCA (Figuras 4 y 6). Aunque por defecto, la distancia de conectividad suele ser de cuatro, este valor depende del número de aminoácidos que se tengan por vuelta. La segunda aplicación, consiste en la facilidad de ubicar las posiciones de ciertos aminoácido dentro de un mismo agrupamiento hidrofóbico; para hacerlo, primero denotamos un aminoácido de referencia, al que llamaremos *"i"*; una vez hecho esto, podemos referirnos a otros que le rodean como "i+4", "i-2", "i+8", etc. Esta forma de ubicar aminoácidos es independiente del carácter polar del mismo (8, 38, 42) [Figuras 4 y 6]. La importancia de resaltar la posición de determinados aminoácidos resulta evidente cuando notamos que los diagramas de HCA presentan un código único que facilita gráficamente la inspección y análisis de algunos aminoácidos (mayoritariamente anfipáticos) como son la Serina, Treonina y Cisteína (Figura 6 y Tabla 2). A pesar de que con este tipo de representaciones podemos resumir en unas cuantas líneas mucha información, las notorias desventajas de las representaciones con distancia de conectividad radican en que son poco amigables al ojo humano (si se maneja dentro de un Diagrama de HCA es difícil comprender el significado *a priori* de las distancias, por lo que, con frecuencia, los análisis de los diagrama de HCA suelen limitarse a la identificación de agrupamientos hidrofóbicos distintos, ignorando, en términos generales, la distancia relativa a la que éstos puedan encontrarse (Figura 4).

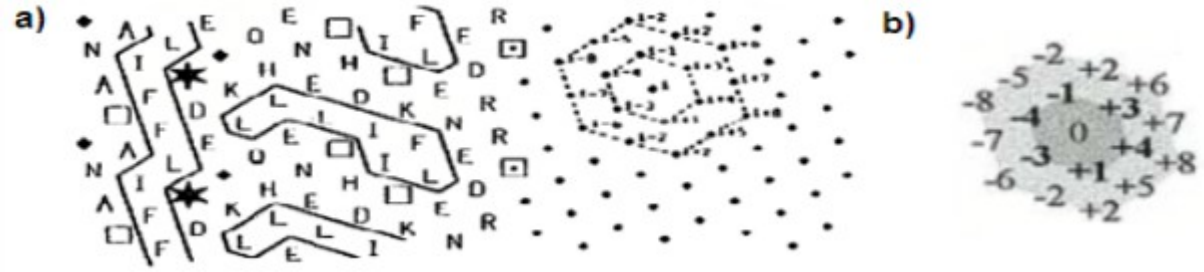

Figura 4. (a) Representación de la distancia de conectividad. Nótese que cualquier posición (aminoácido) en un diagrama de HCA puede ser ubicado y descrito en base a este principio. **(b)** Aquí, el aminoácido central, denominado "0"; es descrito, junto con los aminoácidos que le rodean, en términos de distancia. El hexágono interno (en tono más oscuro) considera el número de aminoácidos que hay en la periferia del aminoácido "0" (al centro); mientras que lo mismo ocurre con el hexágono más externo (tono más claro) que considera el doble de la distancia (aprox. 3.6 aminoácidos más) partiendo del mismo punto. En conjunto, éste diagrama representa las distancias de conectividad entre aminoácidos hidrofóbicos que nos ayudan a determinar cuándo pertenecen al mismo agrupamiento (*cluster*), y la identidad (polar o no polar) de los aminoácidos que le rodean; pues si un aminoácido polar se interpone en un agrupamiento, éste es interrumpido y, a partir de tal punto, los siguientes aminoácidos hidrofóbicos pertenecen a un agrupamiento distinto. Modificado de las Referencias 6 y 8.

Podemos observar que categorizando a los aminoácidos en hidrofóbicos e hidrofílicos es factible expresar la distancia de los agrupamientos hidrofóbicos individuales en código binario, en el que los aminoácidos hidrofóbicos adquieren el valor de uno, independientemente de cuál de ellos sea, mientras que aquellos aminoácidos hidrofílicos adquieren el valor de cero (6, 5, 8, 42), [Figura 5]. Utilizando la distancia de conectividad y/o la naturaleza polar de los aminoácidos que rodea cada agrupamiento, sería relativamente sencillo definir a los agrupamientos hidrifóbicos en términos más precisos que denoten algunas de sus características primordiales [Figura 4]. Codificando a los aminoácidos como ceros y unos con base en su carácter polar (los neutros también se codificarían como “cero”) resulta conveniente; pues, una vez identificados, se facilita trabajarlos en bases de datos o manipularlos de formas complejas (computacionalmente intensivas); además de contar con la ventaja de que se vuelve factible describir no sólo la posición del aminoácido, sino también las características de aquellos que le rodean. El código binario nos permite expresar estructuras secundarias comunes de forma sencilla; por ejemplo, el agrupamiento 10101 (recordemos que los agrupamientos siempre comienzan y terminan con un aminoácido hidrofóbico) podría representar una alfa-hélice anfipática, en las que hay sucesión y alternancia de aminoácidos polares y no polares. En formas más computacionalmente amigables, los agrupamientos hidrofóbicos se expresan en función del valor de P.

En el caso del agrupamiento: 101011 el valor de P equivale a 53. El valor se obtiene del cálculo: ((1x1)+(2x0)+(4x1)+(8x0)+(16x1)+(32x1)) [6, 42].

Aunque útil, esta transformación se debe usar con cuidado, pues implica la pérdida de la información individual que cada aminoácido podría aportar para entender el plegamiento tridimensional de una proteína. Bajo este planteamiento, por ejemplo, una Serina o una Treonina se convierten en cero [Figura 5], lo que implica la pérdida de la información extra que estos aminoácidos proveen cuando se analiza la estructura en un diagrama de HCA de cualquier secuencia. El mismo efecto pero con resultados más dramáticos sucedería si solamente evaluamos una Prolina o una Cisteína en términos de “ceros y unos”, respectivamente (38).

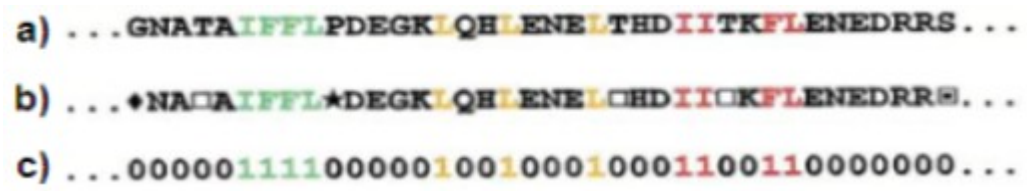

Figura 5. Representación de una secuencia de aminoácidos en código binario para su análisis mediante el método de HCA. Partiendo de una secuencia o segmento cualquiera de ella (a), es posible codificarla con los principios para representar aminoácidos fundamentales dentro del método de HCA. (b) En ocasiones, la representación de secuencias muy largas se puede hacer de una forma más resumida, como lo sería la transformación de la misma secuencia en código binario (c), en donde 1= a.a. Hidrofóbico y 0= a.a NO Hidrofóbico.

Una de las características distintivas de los diagramas de HCA (que como mencionamos son la unidad de análisis del método) es que en ellos se muestra la distribución espacial relativa (codificada a manera de símbolos para cada aminoácido en especial [Figuras 3, 5 y 6, Tabla 2] de todos los aminoácidos que componen una secuencia proteica; sin embargo, se hace énfasis en la distinción de aminoácidos hidrofóbicos que con frecuencia se encuentran formando el núcleo o *core* de las proteínas, aminoácidos polares de importancia estructural y la asociación de los aminoácidos hidrofóbicos en agrupamientos (*clusters*), los cuales podrían estar reflejando la ubicación de estructuras secundarias regulares, como Alfa-hélices, Hojas Beta (Tabla 1 y 2).

La correlación matemática que existe entre los agrupamientos de aminoácidos hidrofóbicos y la predisposición que tienen a formar determinadas estructuras secundarias no debe mal interpretarse (de suma importancia cuando se analizan los diagramas de HCA); pues lo que esto significa es que si un aminoácido hidrofóbico se encuentra con una mayor frecuencia en regiones internas de la estructura (tal vez participando en la formación del core proteico), esto no excluye la posibilidad de encontrar otro del mismo carácter polar en otra región, siendo participe, por ejemplo, de una “asa” (*loop)*. Todos los aminoácidos pueden encontrarse formando diversas estructuras secundarias, independientemente de que, por sus inherentes propiedades físico-químicas, se tienda a favorecer unas de otras (Tabla 1 y 2).

Tabla 1. En esta tabla se muestra la frecuencia de ocurrencia de los aminoácidos dependiendo del tipo de estructura secundaria que tienden a formar. Modificaciones hechas a la imagen de la Referencia 38, página 18.

Amino ácido	α - hélice	Hoja - β	*Loop*	Amino ácido	α - hélice	Hoja - β	*Loop*	Amino ácido	α - hélice	Hoja - β	*Loop*
Glu	**1.59**	0.52	1.01	Val	0.90	**1.87**	0.41	Gly	0.43	0.58	**1.77**
Ala	**1.41**	0.72	0.82	Ile	1.09	**1.67**	0.47	Asn	0.76	0.48	**1.34**
Leu	**1.34**	1.22	0.57	Tyr	0.74	**1.45**	0.76	Pro	0.34	0.31	**1.32**
Met	**1.30**	1.14	0.52	Cys	0.66	**1.40**	0.54	Ser	0.57	0.96	**1.22**
Gln	**1.27**	0.98	0.84	Trp	1.02	**1.35**	0.65	Asp	0.99	0.39	**1.24**
Lys	**1.23**	0.69	1.07	Phe	1.16	**1.33**	0.59				
Arg	**1.21**	0.84	0.90	Thr	0.76	**1.17**	0.90				
His	**1.05**	0.80	0.81								

2.8 Representaciones de los aminoácidos en los diagramas de HCA

En la parte introductoria de éste trabajo ya señalamos propiedades físico-químicas de algunos de los 20 aminoácidos. Sin embargo, en esta sección, analizaremos con más detalle los aminoácidos que tienen una importancia y distinción especial para el método de HCA y los diagramas, pues resulta muy evidente que, cuando observamos un diagrama de HCA como el de la Figura 5, podemos notar que los aminoácidos, simbolizados por su identificador de una sola letra, aparecen coloreados de distintas formas, y en ocasiones denotados con una figura geométrica. Con la intención de facilitar la inspección rápida de los diagramas de HCA, a continuación damos información concreta que permita su interpretación general.

Prolina: Dependiendo de algunos de los perfiles de hidrofobicidad más comunes (60), es un aminoácido que tiende hacia la hidrofobicidad según el pH en el que se encuentre (38). Aunque la Prolina en sí no es hidrofóbica, su estructura cerrada no es propensa a formar enlaces con otras moléculas, por lo que en el método de HCA se le considera como el único aminoácido polar que posee una representación gráfica distintiva en los diagramas (38, 60). Es un aminoácido peculiar por su estructura en forma de anillo voluminoso que le incapacita para formar enlaces con otros aminoácidos, pues el NH2 que los efectúa, se encuentra ocupado encerrando el anillo propio del aminoácido. La forma especial de la estructura de la Prolina (con el anillo llamado imino) impone fuertes limitaciones al plegamiento, por lo que es considerado como un aminoácido capaz de romper agrupamientos (hidrofóbicos e hidrofílicos). Esta característica, que se resalta en la Tabla 1, parece estar asociada a la baja presencia del aminoácido Prolina en estructuras secundarias; sin embargo, este aminoácido suele ocupar tales espacios, y aun no se encuentra un factor común que favorezca dicha presencia. La Prolina se representa gráficamente con estrellas rojas.

Glicina: Es un aminoácido ligeramente hidrofóbico; es el más pequeño y uno de los más versátiles, pues puede mimetizar a otros aminoácidos hidrofóbicos (38). Se representa con estructuras en forma de diamante de color negro.

Serina: Es un aminoácido polar neutro que puede enmascarar su carácter polar llevando a cabo enlaces tipo hidrógeno con el grupo carboxilo. Por su capacidad, suele formar alfa-hélices. Las Serinas se denotan con un cuadro encerrando un punto negro.

Treonina: Es un aminoácido polar y neutro similar en características a la Serina, que sólo se diferencia en su grupo CH3 “extra” en su estructura. Es capaz de mimetizar a la Valina y/o Isoleucina, que son dos de los aminoácidos más hidrofóbicos. Esta característica se sospecha es debida a su grupo metilo de la cadena lateral (8, 38). Las Treoninas se representan con un cuadrado vacío.

Cisteína: Otro aminoácido polar neutro. Es un aminoácido mimético que se puede encontrar en agrupamientos hidrofóbicos o hidrofílicos. La Cisteína es reconocida por su capacidad para formar puentes disulfuro dobles de gran fortaleza. En proteínas extracelulares -que suelen funcionar en ambientes no reductores-, los enlaces disulfuro que puede formar gracias a su grupo tiol, suelen ser primordiales en delimitar el plegamiento de tales proteínas. La Cisteína se denota con una C. Ver Figura 6 y Tabla 2 para más información sobre la representación de éstos aminoácidos.

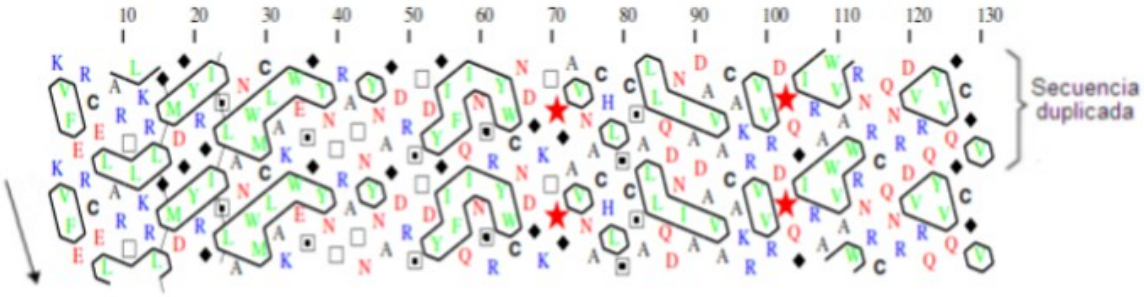

Figura 6. Representación bidimensional mediante los diagramas de HCA de la estructura de la Lisozima tipo C de humano. Los agrupamientos (*clusters*) de aminoácidos hidrofóbicos (letras de color verde) se delimitan y producen figuras características que se pueden asociar a la formación de estructuras secundarias regulares características. La línea descendente muestra el sentido de lectura de la secuencia. Figura Producida con el programa HCA-Draw. Indicadores modificados de la Referencia 5.

Aunque no se denota gráficamente en los diagramas de HCA, los aminoácidos hidrofóbicos importantes para el método: VILFMYW de Valina, Isoleucina, Leucina, Fenilalanina, Metionina, Tirosina y Triptófano se dividen en hidrofóbicos fuertes (VILF) y débiles (MYW), aunque todos se denotan de color verde. Esta división resulta importante en la interpretación de los diagramas, pues el núcleo proteico suele estar conformado, en su mayoría, por los aminoácidos hidrofóbicos más fuertes, mientras que los más débiles pueden presentarse en regiones más expuestas. Además, en el código de HCA no se consideran a todos los aminoácidos hidrofóbicos. Por ejemplo, faltan la Alanina y la Glicina. Estos aminoácidos no están dentro del alfabeto hidrofóbico pues frecuentemente se encuentran formando "asas" o *loops* en la estructura de las proteínas. Estas estructuras irregulares suelen participar en interacciones polares con otros aminoácidos, y como ya hemos mencionado, no se ha podido asociar su presencia con algún tipo de estructura secundaria o tridimensional regular. En el mismo sentido, la Prolina, aminoácido mimético que suele ser hidrofílico a pH 2 (73) tampoco se incluye dentro del alfabeto hidrofóbico del método de HCA; sin embargo, por sus peculiares características, se le considera como un aminoácido de suma importancia pues es capaz de romper los agrupamientos de aminoácidos hidrofóbicos.

Tabla 2. Características generales resumidas de los aminoácidos importantes en el método de HCA. Adaptado de la Referencia 73. Capítulo 2. Páginas: 20-23.

Aminoácido	**Representación Gráfica/Color**	**Caracter Polar a pH 7**	**Características**
Glicina	Rombo Negro	Hidrofóbico	Es el aa más pequeño con sólo un átomo de hidrógeno como cadena lateral, por lo que es muy versátil y se encuentra con frecuencia formando *loops* y alfa-hélices. No es parte del alfabeto hidrofóbico de HCA.
Alanina	**"A"** Negra	Hidrofóbico	Es mimético y puede sustituir aas hidrofóbicos. Suele estar en regiones desordenadas o *loops*. No se considera en el alfabeto hidrofóbico de HCA.
Prolina	Asterisco * Rojo	Hidrofóbico	Es mimético y posee una cadena alifática similar a la de otros aminoácidos hidrofóbicos. Generalmente rompe agrupamientos de hidrofóbicos e impone fuertes restricciones estéricas.
Serina	Cuadrado con punto Negro	Polar Neutro	Es pequeño, mimético y su cadena es alifática. Posee un grupo hidroxilo muy polar que puede servir como sitio de modificaciones post-traduccionales. Suele estar en alfa-hélices.

Treonina	Cuadrado sin punto Negro	Polar Neutro	Es muy similar a la Serina y se intercambia con frecuencia. También posee un grupo hidroxilo muy polar. Puede sustituir aa´s hidrofóbicos como Valina e Isoleucina. Las sustituciones por este aminoácido son muy raras.
Cisteína	**"C"** Negra	Polar	Posee un grupo metilo –CH2- hidrofóbico al igual que un grupo tiol con un átomo de azufre. Este le permite llevar a cabo enlaces disulfuro, dobles o sencillos de gran fortaleza. El grupo tiol es altamente reactivo, cualidad que se aprovecha para llevar a cabo reacciones enzimáticas. Además, este grupo le permite formar enlaces de alta afinidad con metales (ej. Cobre, hierro, zinc, etc.)
Lisina	**"K"** Azul	Polar	Su cadena lateral es hidrofóbica y consta de cuatro grupos metileno; sin embargo, en condiciones acuosas el amino-epsilón se protona y se convierte en muy polar. Suele encontrarse formando alfa-hélices y *loops* en las proteínas. Puede mimetizar a ciertos aminoácidos hidrofóbicos y alternarse con éstos en zonas anfifílicas. Es el aminoácido que más experimenta múltiples modificaciones post-traduccionales.
Arginina	**"R"** Azul	Polar	Su cadena lateral es hidrofóbica y consta de tres grupos metileno y un grupo delta-guanidino altamente básico que, en condiciones acuosas se protona y se convierte en muy polar. Suele encontrarse formando alfa-hélices y *loops* en las proteínas. Independientemente de que poseen estructuras disímiles, suele sustituir a la Tirosina o al Triptófano. Es el segundo aminoácido que más experimenta modificaciones post-traduccionales.

---

3. Las Lisozimas

Las lisozimas son proteínas modelo en el estudio estructural de las biomoléculas cuyas propiedades enzimáticas están bien estudiadas y definidas. Son enzimas que pueden romper el enlace glucosídico entre dos o más péptidos; esto, lo consiguen hidrolizando el enlace 1,4-beta entre el ácido-acetil-murámico y el ácido-acetil-glucosamina (16, 17, 20). La gran variedad de sustratos en los que las Lisozimas pueden actuar ha favorecido el desarrollo de estudios cristalográficos en los que se ha analizado el sitio catalítico (66, 67). Existe evidencia de que en la estructura de las Lisozimas, sólo los aminoácidos Asparagina (que es el donador de protones) y Acido Glutámico (nucleófilo) están muy conservados, al igual que las estructuras en forma de

hojas-beta plegadas en las que están embebidos. Trabajos posteriores han identificado otros aminoácidos (mediante ensayos de mutagénesis dirigida) que podrían ser importantes para la actividad catalítica y que se encuentran predominantemente en Lisozimas de otros organismos u otros tipos (66, 67, 73).

Aunque se les reconoce principalmente por su capacidad anti-microbial, pues pueden lisar bacterias, las Liozimas poseen otras funciones interesantes, como: la hidrólisis de polisacáridos de almacenamiento y también se emplean como proteínas bacteriotóxicas.

Todas estas enzimas se clasifican dentro de las Glucósido-Hidrolasas (45, 55, 56, 61, 62, 66), aunque no comparten exactamente la misma clasificación en todas las bases de datos. La Lisozima Tipo C (revisada en detalle más adelante) es la más estudiada a diversos niveles.

Se piensa que las Lisozimas están presentes en todos los reinos de organismos (1, 17); sin embargo, por su secuencia divergente, en ocasiones no han podido ser adecuadamente identificadas empleando métodos de búsqueda tradicionales, pues alinear correctamente las secuencias de éstas enzimas puede resultar complicado. Herramientas comúnmente utilizadas como P-BLAST con frecuencia no pueden identificar correctamente todas -o al menos la mayoría de las secuencias idénticas-, presumiblemente homólogas de éstas proteínas, por lo que los análisis filogenéticos que se producen empleando las secuencias identificadas suelen ser restringidos e incompletos y usualmente enfocados a Eucariontes multicelulares, así que, en comparación, se conoce relativamente poco de las Lisozimas procariontes, de insectos y hongos.

La tasa de variación de los genes de las Lisozimas ha producido una variedad notable de enzimas, esto favorece que identificarlas, analizarlas y categorizarlas se vuelven labores complejas. De hecho, Irwin y Biegel (16) al hacer búsquedas en la base de datos ENSEMBL (63) para encontrar secuencias homólogas de lisozimas en vertebrados, se percataron que algunos organismos tenían más de una copia de ciertos genes (p. ej. de LyzL1). Para demostrar que tales copias eran resultado de eventos de duplicación, hicieron un análisis de “vecinos genómicos” [*Genomic Neighborhoods*]; ya que la variabilidad y versatilidad de las Lisozimas favorecen fenómenos como el de reemplazo de genes no-ortólogos, además de sustituciones ortólogas y duplicaciones; lo que dificulta mucho la labor de separar genes parálogos de ortólogos que se puedan utilizar para las reconstrucciones filogenéticas más confiables (16).

Considerando varios de los obstáculos que se necesitan superar en el análisis de enzimas con este tipo de características, el conocimiento que se tiene de las Lisozimas proviene principalmente de

estudios estructurales, de caracterización, de aislamiento, de mutagénesis y del análisis *in silico* de las secuencias depositadas en las bases de datos. Como resultado directo de su estudio detallado y dado su nivel de variación, las glucósido-hidrolasas se clasifican en más de 100 familias (23, 55); aunque, Lisozimas como tal, sólo se reconocen alrededor de 5 tipos (17). De acuerdo a la clasificación de hidrolasas que se hace en UniProt [61] y HOMSTRAD [55], en este trabajo nos enfocamos en representantes de la denominada Familia 22. Aquí manejaremos una nomenclatura similar a la que encontramos en la base de datos SCOP (62); allí se clasifican como: Lisozimas tipo C, G, Muramidasas con actividad catalítica, Glucosidasas de la Familia 19 [una división de las 100 que existen para la súper-familia (23)] y Lisozimas de fagos, (Lambda, P1 y T4, por nombrar algunos). En la misma clasificación se reconocen otras familias, pero realmente ya no se les considera como Lisozimas en sentido estricto, pues en estos casos los análisis son escasos y sólo mediante ensayos de actividad catalítica y métodos comparativos (como alineaciones de perfiles con BLAST) se asume que la similitud de secuencia.

Considerando lo anterior, detallaré un poco sobre las categorías de Lisozimas que aparecen en la base de datos de SCOP (62) y CATH (45).

Lisozima tipo C: Es la lisozima "canónica" y más estudiada. Recibe su nombre de la mano de Alexander Flemming, quien la descubrió en la clara de huevo (la letra "C" denota que se descubrió en el pollo) en 1923. Como Flemming observó que la enzima era capaz de lisar bacterias, fueron descritas como Lisozimas. Las lisozimas tipo C cuentan con una amplia pero ligeramente sesgada distribución filogenética, pues además de estar presentes en los mamíferos, también se han encontrado en algunos insectos, reptiles, peces y artrópodos (17). Las filogenias que se han obtenido utilizando esta enzima como marcador, ponen de manifiesto la tasa acelerada de mutación para el gen, pues la divergencia a nivel de secuencia es muy variable, y en los casos en los que se ha intentado construir filogenias utilizado éstos genes como marcadores moleculares, se han reportado numerosas agrupaciones que resultan muy incongruentes cuando se comparan con una filogenia de especies (16, 17, 19). Uno de los aspectos más estudiados de las Lisozimas tipo C es que pudieron ser precursoras de otro tipo de Lisozimas que, aparentemente, "modificaron" la función original, que era romper peptidoglicanos. Cuando la enzima ya no se encontró bajo la misma presión de selección adquirió una nueva función, que es unir calcio. Esta enzima, llamada Alfa-lacto-albúmina (LALBA) es descendiente de la Lisozima tipo C (16, 29, 30, 49). Estudios comparativos/filogenéticos han demostrado que existe otro tipo de Lisozimas que

descienden de las Tipo C. Estas Lisozimas, también pueden unir calcio pero no han perdido la función original (en este trabajo nos referimos a ellas como Lisozimas CB) [16, 49, 73].

Basados en la identificación de la estructura intrón-exón en los genes de Lisozimas y en análisis de vecinos genómicos (*Genomic Neighborhoods*); en el trabajo de Irwin y Biegel [16] se propone que hay, al menos, ocho genes de Lisozimas en el genoma de varios mamíferos; sin embargo, muchos de ésos genes muestran una aparente neo-funcionalización hacia papeles más específicos, algunos de los cuales podrían ser especie-específicos, como las Lisozimas con papel reproductivo (16). La divergencia de la alfa-lactoalbúmina es de mucho interés, pues al parecer, la enzima se encuentra sólo en el linaje de los mamíferos, así que se postula que la neo-funcionalización de la Lisozima tipo C que dio origen a las alfa-lactoalbúminas ocurre después de la separación que se da entre las aves y los mamíferos (16). Esto suele considerarse como plausible dado que, al contrario de los mamíferos, ningún ave tiene alfa-lactoalbúmina, muy pocas especies de mamíferos y aves tienen Lisozimas CB (pueden unir calcio, pero que no es su función primaria) pero si poseen copias de Lisozimas tipo C y G.

Podemos notar que el origen y la distribución filogenética de las Lisozimas tipo C, la alfa-lactoalbúmina y las Lisozimas CB es un tema interesante, aunque presenta retos importantes, dado que es complicado identificar sólo con la secuencia los genes que podrían ser homólogos. Aunque en este caso podrían estar ocurriendo reemplazos de genes ortólogos (38, 73), no hay evidencia suficiente para apoyar la hipótesis, y es así que esperamos que con trabajos como el de nosotros, a futuro sea plausible la identificación -en un genoma completo- de genes/segmentos ortólogos independientemente del nivel de divergencia en la secuencia; en cuyo caso, para obtener resultados confiables, es imperativo evaluar la estructura. Al mismo tiempo, hipotetizamos que el método podría aportar evidencia a favor o en contra de la similitud estructural debido a un origen común de todas éstas enzimas. De lograrse, es muy factible que se vaya cerrando una discusión que lleva abierta más de 25 años, en donde la carencia de secuencias/estructuras, la imposibilidad de los análisis a gran escala y la dificultad de las evaluaciones precisas, han impedido llegar a una conclusión.

Lisozima tipo G: Se encuentra principalmente en aves. Aparentemente, fue descubierta en los gansos, pues la familia lleva como nombre "*Goose Type Lysozyme*" (17). En algunas aves, predomina la presencia de la Lisozima tipo G, mientras que en otras es la de tipo C (principalmente

en en representantes del orden *Galliformes*) y en otras tantas podemos encontrar ambos tipos. Lo más interesante, es que ambos tipos de enzima se encuentran fuera de la clase de las aves. Mediante análisis filogenéticos y búsquedas mediante BLAST, se ha puesto de manifiesto que este tipo de lisozimas tienen representantes en reptiles, peces, mamíferos, algunos moluscos (invertebrados). En el trabajo de Callewaert y Michiels (17) se propone que se tratan de secuencias funcionales. De igual manera, en el mismo trabajo hacen notar que no encuentran secuencias ortólogas en organismos como *Drosophila melanogaster* y *Caenorhabditis briggsae*; sin embargo, como hemos hecho notar mucho en este trabajo, las secuencias son muy divergentes (aun entre las Lisozimas del mismo tipo) y por tanto es posible que programas como BLAST, que se basa en los análisis de la secuencia, sea incapaz de detectar homólogos distantes de tales enzimas.

Lisozima tipo "i": Es la que se ha encontrado en organismos invertebrados (de ahí su nombre); anélidos, principalmente, pero que incluye equinodermos, moluscos y algunos artrópodos (16, 17). Se propuso la existencia de éstas Lisozimas comparando la región N-terminal de la enzima con las versiones que tienen otros organismos (ver referencias dentro de 17 para más detalles); pero fue sólo hasta 1999 cuando se descubrió por primera vez en la estrella de mar *Asterias rubens* (17). Las primeras evidencias que confirmaban que se trataba de Lisozimas tipo "i" provenían de la confirmación de la actividad catalítica; aunque ahora sabemos que esto puede ser "tendencioso", pues muchas Lisozimas tienen la actividad independientemente del tipo. Callewaert y Michiels (17) proponen que se han encontrado Lisozimas de tipo "i" en prácticamente todos los insectos estudiados; aunque, de forma simultánea y de forma muy interesante, se ha reportado que no se encuentran genes ortólogos en muchos vertebrados, como *Bos taurus*, *Canis lupus familiaris*, *Equus caballus*, *Homo sapiens*, *Mus musculus*, etc. ¿Será que este tipo de Lisozimas cuentan una historia evolutiva distinta?. ¿La divergencia de la secuencia nuevamente limita los resultados?. Hasta el momento, no hay más estudios al respecto.

Lisozimas de Fagos: Es una categoría que sólo se encuentra en las bases de datos de SCOP, HOMSTRAD y en el portal CAZYPEDIA (23), en donde, de hecho, se encuentran divididas en Lisozimas de Fago Lambda y T4 (en HOMSTRAD se reportan incluso las de otros fagos). En otras bases de datos y en la literatura suelen categorizarse como Lisozimas Tipo V (provenientes de virus); sin embargo, no todas se clasifican dentro de la misma categoría, principalmente porque, a nivel de secuencia, existen muchas dudas sobre un posible origen común como los otros tipos de Lisozimas. Todas las Lisozimas que se han encontrado en Fagos que muestran similitud con las

Lisozimas de Bacterias y viceversa (debido, posiblemente, a eventos de re-inserción) se clasifican dentro de la familia 24 de las glucósido hidrolasas (61). Las primeras secuencias analizadas de virus que fueron descubiertas pertenecen al Fago T4; aunque, posteriormente, se encontraron más en otros virus, como el fago Lambda y el P21 (16, 17, 27, 31, 53, 54).

Lisozimas de Bacterias: A excepción de la Pesticina, que es la enzima con función catalítica y similitud estructural a otros tipos de Lisozimas que mejor se ha estudiado (54), existe muy poca información para este tipo de secuencias. La distribución filogenética sesgada, la falta de estructuras resueltas para comparar, la divergencia de la secuencia y la modularidad de la estructura proteica (con componentes individuales de varias Lisozimas se pueden construir variantes quiméricas funcionales) aunado a que algunas enzimas han sufrido duplicaciones que eventualmente han producido variantes neo-funcionales que poseen una gran diversidad de substratos y en los que además varía la afinidad, han dificultado una interpretación certera y confiable de cómo y dónde surgieron éstas enzimas (17, 31, 53, 54).

El descubrimiento de la similitud estructural entre las Lisozimas de Bacterias, Fagos y Eucariontes apoyan un origen común, pero del cual se desconocen muchos detalles (49, 53, 54, 66, 67). Todas las Lisozimas encontradas poseen actividad catalítica (aunque variable) y se ha demostrado que son funcionales y de utilidad para el organismo que las produce. En pruebas realizadas de forma exclusiva para este trabajo, se hicieron múltiples corridas de HMMER y PSI-BLAST de las versiones encontradas en el fago T4 y Lambda. Los resultados arrojan que hay similitud con las Lisozimas de *Escherichia coli*, *E. Fergusonii*, *Salmonella typhimurium* y *Shigella flexneri*; así como con algunas otras Enterobacterias, que por su frecuencia de aparición baja no mencionaré. Algo interesante es que el HMMER del Fago Lambda arroja como resultado positivo a la Pesticina (con identidad del 85% y un *valor de E* muy bajo de 5e-95), que es la variante que posee *Yersinia pestis*. Esta enzima posee un domino muy conservado entre las Lisozimas; sin embargo, ha adquirido una mayor flexibilidad estructural (se puede desnaturalizar y re-naturalizar) así como todo un dominio nuevo que le facilita la interacción con proteínas y canales transportadores de membrana (54). Esta característica, que es intrínseca de proteínas de defensa celular que pertenecen a otro grupo de proteínas de Bacterias Gram-Negativas (bacteriocinas), pone de manifiesto la versatilidad de las Lisozimas y los múltiples eventos de diversificación que posiblemente han favorecido su persistencia (54).

Además de las categorías arriba mencionadas, se han encontrado secuencias repetidas y muy símiles en los genomas completos de hongos (17) e insectos (Lisozimas tipo “i”), en algunos crustáceos y hasta en cnidarios (16, 17). Nosotros, además, empleando HMMER, encontramos resultados que coinciden con varias Lisozimas-L de Arqueas (que han sido muy poco estudiadas). La baja cantidad de reportes enfocados en el estudio (evolutivo o funcional) de estas enzimas no ha favorecido la resolución definitiva de cuál y cómo ha sido su origen.

Si bien es cierto que todas las secuencias de las Lisozimas muestran algún tipo de similitud (principalmente estructural), esta es, en general, altamente divergente, y aunado a la distribución filogenética aparentemente limitada de algunos tipos (P. ej. Tipo “i” y G) se ha generado cierta incertidumbre respecto a si todas estas enzimas son en realidad homólogas o son el producto de eventos de reemplazo de genes ortólogos y transferencia horizontal. Es posible que éstos fenómenos resulten más evidentes en algunos casos, como en aquellos organismos que intercambian material genético de forma frecuente (como son las bacterias y ciertos eucariontes parásitos) o en los que poseen un genoma muy plástico; lo que, ciertamente, no explica satisfactoriamente los patrones de ocurrencia y variabilidad que observamos en las Lisozimas (26, 27, 29, 30). La contraparte propone que muy seguramente son homólogas, aunque es el nivel de divergencia de la estructura primaria (secuencia) lo que oscurece y dificulta la resolución del posible origen común de todas estas enzimas (17, 18, 19, 28, 54, 66).

4. Justificación

Como se mencionaba en la introducción de este trabajo, el incremento en la capacidad de secuenciar genomas completos y la acumulación progresiva y constante de secuencias en las bases de datos, exige de un método eficiente, rápido y confiable capaz de analizar toda ésa información. Los métodos tradicionales poseen requisitos que en la mayoría de los casos no se pueden cumplir, y en los casos en los que es posible contar con ellos, el tiempo (y a veces costo) de los análisis es prohibitivo. Además, hay que tomar en cuenta que con frecuencia tales metodologías podrían no ser concluyentes. Por ahora, las bases de datos contienen, en su mayoría (aunque ciertamente existen aquellas con información estructural detallada, como el PDB (22) y SCOP [23]) secuencias de nucleótidos/aminoácidos de las cuales se puede extraer información sumamente valiosa; el reto, ahora, es desarrollar y/o aplicar metodologías con sustentos teóricos robustos que nos permitan utilizar este tipo de datos sin la necesidad de hacer evaluaciones complejas y/o muy

tardadas, con un margen de error aceptable (o por lo menos cuantificable), capaces de limitar el sesgo y que nos permitan analizar grandes cantidades de datos.

Nosotros proponemos usar, bajo el marco teórico del método de HCA, el algoritmo HCA_Analyze_Multiple_Aligns (HCAmA) para extraer información directamente de las secuencias de las Lisozimas de Bacterias. Dada la sencillez (sólo se requiere de la secuencia de aminoácidos de una proteína cualquiera para hacer un diagrama de HCA) y plasticidad del método (ya se cuenta con una forma de discriminar y evaluar dentro de un gran número de secuencias, obteniendo una medida de incertidumbre en los resultados) creemos que tiene un enorme potencial para trabajarse en casos donde la estructura de una biomolécula es de interés primordial para resolver cualquier tipo de hipótesis planteada.
Aquí consideramos que metodologías que sólo requieran de una secuencia para analizar y con ella poder hacer inferencias directas (de tipo predictivo) o comparadas (entre varias secuencias) son el futuro para hacer un acercamiento fundamental a favor de un entendimiento más íntegro de la información que tenemos en nuestras bases de datos, y que, obviamente, se vería reflejado en el aumento de la capacidad para entender diversos aspectos de los organismos de los cuales las obtenemos, como la Ecología, la Microbiología, la Genética, la Sistemática y la Evolución.

El algoritmo HCA_Analyze_Multiple_Aligns (HCAmA) nos permite hacer evaluaciones rápidas de una gran cantidad de secuencias, además de que se ha confirmado que es confiable y de gran utilidad en el estudio de secuencias con un nivel de divergencia notable (<25%) [5]. Por esta razón, decidimos enfocarnos en las Lisozimas; ya que representan una oportunidad de estudio único para nuestros objetivos, pues su gran diversidad, su modularidad (poseen dominios “intercambiados”), nivel de divergencia y su predominancia (en una escala filogenética), las convierten en un buen candidato para analizarlas bajo una metodología que nos permita superar las limitaciones técnicas/prácticas y hacer estudios a gran escala y, de igual forma, aplicar evaluaciones simultáneas a los trabajos de análisis de secuencia y/o estructura (normalmente llevados a cabo de forma separada) para favorecer y maximizar las posibilidades de obtener un resultado confiable.

Entre las Lisozimas más similares, segmentos conservados en la secuencia que indican la posible formación de estructuras secundarias regulares (alfa-hélices y hojas-beta) así como un plegamiento idéntico, son fuertes indicadores de que las enzimas son homólogas; sin embargo,

nosotros decidimos enfocarnos en aquellas que son más disímiles (>30% identidad), es decir, las Lisozimas de Bacterias. Estas enzimas suelen ser altamente divergentes, y aunque la gran mayoría no cuentan con estructura tridimensional resuelta, todavía se consideran como Lisozimas, pues se trata de enzimas que son agrupadas dentro de una categoría que reúne a todas aquellas que pueden realizar funciones características de las Lisozimas. Nosotros tomaremos el reto de analizar una muestra (*dataset*) de más de 100 secuencias (algo que, hasta donde llega el conocimiento de este autor, nunca se ha hecho), con lo que se pretende cumplir los siguientes objetivos.

4.1 Hipótesis

I) Basándonos en los trabajos que se encuentran en la literatura, aquí partimos de la hipótesis de que todas las Lisozimas poseen un origen común; sin embargo, por limitaciones en cuanto a la disponibilidad de la estructura tridimensional resuelta de tales enzimas para su comparación, este hecho todavía no ha podido ser confirmado experimentalmente.
II) Hipotetizamos que los Patrones Distintivos de Hidrofobicidad nos permitirán identificar características estructurales comunes entre las secuencias de las Lisozimas bajo estudio, lo que facilitaría hacer inferencias sobre las similitudes estructurales que poseen.

4.2 Objetivos generales

I) Identificación de patrones de hidrofobicidad distintivos en la secuencia de las enzimas Lisozimas mediante evaluación teórica con el algoritmo HCA_Analyze_Multiple_Aligns.

II) Resumir toda la información de las comparaciones estructurales de tal forma que podamos observar gráficamente las características estructurales que comparten las Lisozimas en estudio.

4.3 Objetivos particulares

I) Construir un dendograma estructural en el que podamos mapear los patrones distintivos de hidrofobicidad (PDH) para observar su distribución.

II) Identificar los agrupamientos (clados) que se forman entre las secuencias en las que se encontró al menos un patrón distintivo de hidrofobicidad.

III) Identificar y cuantificar aquellos patrones de hidrofobicidad que encontremos en las Lisozimas de Bacterias.

5. METODOLOGÍA.

Decidimos usar un algoritmo de búsqueda similitud de secuencias que fuese lo suficientemente sensible para encontrar homólogos distantes de nuestra secuencia templado o *Query*. Decidimos emplear HMMER (22), que es un programa de búsqueda de secuencias con una sensibilidad mucho mayor que la alternativa más común, BLAST (13), pues utiliza un algoritmo iterativo basado en los modelos ocultos de Markov para encontrar posibles regiones de similitud aun entre secuencias muy divergentes.

Trabajamos con la versión de HMMER (v3.0 [22]) que se puede accesar desde un servidor dedicado en internet (http://hmmer.janelia.org/) y desde allí hicimos búsquedas en la base de datos del *Protein Data Bank* (PDB, http://www.rcsb.org/pdb/home/home.do) [32] para encontrar coincidencias o "*hits*" de las secuencias en estudio. Una versión resumida de los pasos importantes que seguimos en la metodología se encuentran en el diagrama de la Figura 7.

5.1 El Dendograma estructural

El dendograma (Figura 8) fue hecho en MEGA 6 (69) con el algoritmo de distancia Neighbor-Joining [70] y 5,000 réplicas de *bootstrap*. El dendograma está hecho a escala, en la que las longitudes de las ramas corresponden a la distancia genética, la cual está en términos del número total de sustituciones por sitio. Utilizamos la matriz de JTT (Jones-Taylor-Thornton) como modelo para inferir la distancia genética. Preferimos emplear la matriz de JTT ya que está basada en una gran cantidad de observaciones empíricas (en secuencias divergentes) que toman en cuenta todas las sustituciones esperadas entre los 20 aminoácidos (69). Cabe aclarar que intentamos utilizar la matriz de HCA (descrita en párrafos anteriores) para modelar las distancias genéticas y con ello construir el dendograma, sin embargo, creemos que debido a incompatibilidades con el modelo que describe las características estadísticas del alineamiento (una distribución Gamma, ver más adelante) y a que no establecimos parámetros numéricos para el cálculo de los vectores de la matriz (69), MEGA 6 no nos permitió utilizar la matriz de HCA. A pesar de ello y aun tomando en cuenta que nuestro alineamiento toma un mayor énfasis en los aminoácidos hidrofóbicos, creemos la matriz de JTT podría maximizar la extracción de información útil del mismo.

Sabemos *a priori* que nuestras secuencias son muy variables, así que decidimos modelar la variación entre sitios. Analizamos nuestras secuencias empleando jModelTest 2 (83). Este programa nos estableció que una distribución *Gamma* con un valor de alfa=1 serían los parámetros que mejor describen la variabilidad de las secuencias que trabajamos (69, 83, 84). La distribución *Gamma* con un valor de alfa=1 daría más peso a las modificaciones en las posiciones del alineamiento que cambian poco y otorgaría "libertad" a aquellas que varían demasiado (83, 84), algo que es correspondiente a como se ha observado es la tasa de sustitución para un gen de Lisozima (16, 75). Por el contrario, como deseábamos hacer 5,000 réplicas de *bootstrap*, no quisimos saturar el algoritmo de variables y, dado que nuestras 172 secuencias finales únicas incluían solamente a Bacterias, no consideramos necesario ajustar la variabilidad que se esperaría entre secuencias de organismos pertenecientes a diferentes dominios, por lo que seleccionamos como "uniformes" los patrones de variación entre linajes. Basándonos en las recomendaciones del instructivo del programa MEGA 6, incluir ésta variable en nuestro análisis -cuando de verdad no necesitamos modelar tal fenómeno-, posiblemente sólo aumentaría el tiempo de cómputo; mientras que resulta incierto saber si resultaría benéfico para el resultado (69).

5.2 La Búsqueda Secuencias Homólogas Distantes de la Lisozima Tipo C

En este trabajo buscábamos hacer una aproximación general en la comparación a nivel estructural de diversas Lisozimas. Nuestro objetivo en la delimitación de nuestras secuencias de interés consistió en hacer búsquedas con HMMER utilizando la Lisozima Tipo C de ave (número de acceso en el PDB: 3a8z) como parámetro comparativo (es la secuencia "Templado" o *Query*). Esta secuencia fue seleccionada porque es la que mejor se ha estudiado a detalle en diversos niveles, como estructural, bioquímico (funcional) y evolutivo (elucidando su posible origen común mediante comparaciones con la estructura resuelta de otras Lisozimas disponibles (54, 66, 67).
En la literatura se ha reportado que posee un motivo catalítico en el cual, sólo los aminoácidos que llevan a cabo la catálisis están muy conservados. Este motivo catalítico suele ser distintivo porque se rodea de estructuras secundarias regulares, también muy conservadas, en las que predomina una hoja beta plegada (66, 67). Esta disposición de los dos residuos catalíticos importantes, el nucleófilo y el donador de protones, se encuentra en diversas enzimas de la familia de las glucósido-hidrolasas, inclusive en aquellas que no poseen la actividad catalítica distintiva y que, en ocasiones, pertenecen a otra categoría funcional (54, 66, 67).

Nosotros quisimos utilizar como secuencias templado (*Query*) aquella Lisozima proveniente del evento de divergencia más antiguo; sin embargo, tal cuestión no se puede cumplir dado que el nivel de divergencia y variabilidad entre ellas (por lo menos 30 familias reconocidas de glucósido hidrolasas, [66]) impide la delimitación de aquella secuencia de Lisozima que pudiese cumplir con tales condiciones (considerarse como "ancestral").

A pesar de esto, en la literatura se reporta que la Lisozima Tipo C de Ave es buen candidato para analizar patrones de divergencia de estas enzimas (y de aquellas con características similares) dado la amplia distribución filogenética y a los motivos catalíticos conservados que en ella se encuentran (66). Ya que teníamos claro que secuencia utilizaríamos para hacer las comparaciones, utilizamos *JackHmmer* para hacer las búsquedas de forma iterativa y así encontrar secuencias homólogas distantes.

Es importante mencionar que en HMMER no es aconsejable hacer demasiadas iteraciones, pues el algoritmo, basado en la construcción de perfiles (*profiles*) para realizar búsquedas con los modelos ocultos de Markov, podría comenzar a reportar sólo resultados muy similares entre sí, lo que disminuiría significativamente la diversidad de los organismos seleccionados e incluso la variabilidad de las secuencias consideradas como significativas (22). Aunque nuestro objetivo no es tener una muestra con una gran cantidad de secuencias o muy bien representada, es fundamental procurar que los resultados reportados incluyan secuencias homólogas distantes.

Dicho esto, reportamos que se hicieron cinco iteraciones (para mantener una muestra variable) utilizando la Lisozima Tipo C como secuencias templado (*Query*). La búsqueda se hizo en la base de datos del NCBI con secuencias no-redundantes y evitando aquellas que fueran resultado de análisis ambientales o experimentos de mutagénesis. La penalización a la apertura y extensión de *gaps* se mantuvo en sus valores por defecto; sin embargo, con la intención de maximizar la posibilidad de encontrar secuencias divergentes (<30% de identidad) se utilizó la matriz de BLOSSUM 45 en lugar de la de defecto. Esta matriz es considerada como apropiada para hacer búsquedas de secuencias con homología distante.

5.3 Alineamiento de las Secuencias

Una vez que se tenían las secuencias que se encontraron con HMMER, el siguiente paso fue alinearlas para poder analizarlas con el algoritmo HCAmA. Al igual que con otros métodos que trabajan con la predicción de la estructura de las proteínas o el análisis de dominios y/o motivos

conservados; el método de HCA requiere de un alineamiento entre las secuencias que se trabajarán [38] (sólo la construcción del diagrama de HCA no lo requiere). A pesar de que existe una gran cantidad de programas informáticos para alinear secuencias, por su conveniencia y sencillez se empleó el programa ClustalX (68); pues ClustalX permite de una manera muy accesible cambiar la matriz con la que se guiará el alineamiento. Esta característica es de suma importancia para nuestro trabajo, ya que en literatura se ha reportado lo conveniente que es utilizar una matriz especial para analizar la distribución de los aminoácidos hidrofóbicos como caracter estructural (5, 8, 65, 66), y de esta manera maximizar la información que podemos obtener con este tipo de análisis.

Siendo así, en lugar de la matriz por defecto de ClustalX, se utilizó una matriz que llamamos "la matriz de HCA" (ver más adelante) y que se provee de forma gratuita en el paquete de Pedro Silva (5). Utilizando esta matriz, se alinearon 729 secuencias de Lisozimas de distinta longitud. La más pequeña fue de 88 aminoácidos, mientras que la más grande fue de 942. La media fue de 310 aminoácidos.

Quisiéramos mencionar que las secuencias se alinearon utilizando el total de su longitud; es decir, no fueron editadas en los extremos. Esto se decidió así pues tal procedimiento de edición podría predisponer a la eliminación de aminoácidos importantes para el método (los siete más hidrofóbicos, VILFMYW) lo que podría tener consecuencias poco deseables, como impactar al algoritmo discriminador (habría sesgo en el *HCA_Score*) y dificultar el análisis visual de las secuencias (utilizando los diagramas de HCA) en caso de ser requerido. Además, la longitud total, la proporción y la ubicación de aminoácidos hidrofóbicos son características que el algoritmo correlaciona matemáticamente para determinar aquellas secuencias que están por encima o debajo del umbral de aceptación (significancia), por lo que la deleción de segmentos de la secuencia podría introducir un sesgo cuyo impacto podría ser difícil de establecer con precisión.

5.4 La Matriz de HCA

Dado que el método de HCA trabaja con la comparación de la frecuencia y distribución espacial de los aminoácidos hidrofóbicos que por sus propiedades suelen "juntarse" y formar agregados o "*clusters*", es altamente aconsejable alinear las secuencias en base a éstas mismas características (5, 8, 66).

Dentro de los paquetes de Pedro Silva, se encuentra una matriz especial que en este trabajo denominamos como “Matriz de HCA” (5). Esta matriz sustituye a la que se elije por defecto dentro de ClustalX (14) (PAM o BLOSSUM) y su labor es favorecer el alineamiento de los agrupamientos hidrofóbicos, al igual que residuos de Prolina y Glicina (ver en la descripción del método la importancia relativa de éstos aminoácidos para el método) entre las secuencias. Con esta matriz, se penaliza la sustitución no-sinónima de cualquiera de los aminoácidos hidrofóbicos del conjunto VILFMYW; mientras que cambios sinónimos adquieren un valor unitario. De forma similar, las coincidencias para Prolina y Glicina tienen el mismo valor.

Sustituciones “silenciosas”, como por ejemplo, la que puede ocurrir entre Serina y Treonina, toman un valor de 0.7. Cualquier otro tipo de cambio tiene un valor de cero (5). De esta manera, se le da un peso preferencial a las coincidencias entre aminoácidos hidrofóbicos.
Otra de las cuestiones que me gustaría mencionar, es que la matriz de HCA no es fija y es fácilmente modificable en programas tan comunes como editores de texto estándar, incluyendo programas como Excel. El investigador puede “jugar” con los valores de penalización para las distintas sustituciones de los aminoácidos hidrofóbicos y así “optimizar” el análisis para proteínas de características distintivas. Por ejemplo, si se analiza un grupo de proteínas similares en los que haya una notoria tendencia a sustituir aminoácidos hidrofóbicos por otros miméticos, se puede aumentar o disminuir la penalización (según sea el interés) para ésos cambios y con ello “acomodar” la variabilidad natural que se encuentra en las secuencias de las proteínas, en especial de aquellas regiones que no se encuentran bajo presión de selección o que son naturalmente hipervariables (5, 8, 38, 50, 65).

5.5 Parámetros importantes en el alineamiento

En el artículo de Pedro Silva (5) se menciona que el programa es susceptible a la introducción de *gaps*. Muchos *gaps* pueden favorecer las coincidencias azarosas ente agrupamientos de aminoácidos hidrofóbicos; por ende, se penaliza la apertura de *gaps* con un valor mínimo de tres (5). Esto es variable, y en el trabajo se aconseja que de analizar secuencias más divergentes, se aumente la penalización para la apertura de *gaps*. Se advierte que si no se penaliza apropiadamente la introducción de *gaps* durante el alineamiento, ClustalX podría favorecer coincidencias “azarosas” que impacten negativamente el poder discriminatorio del algoritmo.

Lo anterior, en conjunto con la Matriz de HCA (ver la descripción más adelante), facilita la obtención de un alineamiento en donde se maximiza la identidad entre agrupamientos de aminoácidos hidrofóbicos, lo que puede favorecer que los resultados obtenidos por el algoritmo de HCA_Multiple_Analyze sean más confiables (pues detectan mejor los falsos positivos producto de artefactos metodológicos) [5, 8]. Por tal motivo, en el alineamiento hecho con ClustalX se utilizó una penalización a la apertura de *gaps* de 60. Como nuestras secuencias de Lisozimas son altamente variables, notamos que ciertamente podrían estar "confundiendo" al algoritmo. Cuando esto sucede, *HCA_Analyze* arroja una gran cantidad de resultados "probables" en los que se advierte explícitamente que la cantidad de *gaps* dificulta la diferenciación de resultados significativos de posibles falsos positivos; comportamiento que el usuario puede advertir observando los valores relativamente bajos de *HCA_Score* y/o valor de "E".

Para encontrar un valor de apertura de *gaps* que mejor se ajustara a nuestro alineamiento, hicimos pruebas (los resultados de las mismas no se muestran) con penalizaciones de 30, 40 y 50, antes de decir optar por el valor de 60, en donde el número de resultados "dudosos" disminuía notoriamente cuando se les comparaba con penalizaciones más laxas. Aunque se hizo una prueba con un penalización de 70, observamos que es posible que los datos, así como el algoritmo, podrían estar llegando a un punto de saturación, en donde seguir aumentando el valor de la penalización aparentemente ya no influye en la sensibilidad del algoritmo para identificar correctamente resultados significativos.

5.6 El algoritmo HCA_Analyze_Multiple_Aligns (HCAmA)

En el sitio (http://www2.ufp.pt/-pedros/HCA_db_index.htm), se pueden encontrar los programas, hechos en C, para llevar a cabo estos análisis. Hay dos versiones; la primera, "pareada" (*Pairwise)* (*HCA_Analyze*) sólo hace comparaciones pareadas entre dos secuencias. Esta versión del algoritmo difiere de la otra pues no busca patrones hidrofóbicos distintivos entre las secuencias comparadas, así que por esta razón no lo utilizamos. La segunda versión HCA_Analyze_Multiple_Aligns efectúa los análisis entre múltiples secuencias y en ellas detecta los patrones distintivos de HCA. En ambos casos el código ya está listo para trabajar en Windows (en el caso de Linux hay que compilar el código fuente). Como archivo de entrada, ambos programas utilizan un alineamiento en formato *.pir*.

Dado nuestro interés de evaluar más de un par de secuencias entre sí y de utilizar los patrones distintivos de HCA (PDH), utilizamos el algoritmo HCA_Analyze_Multiple_Aligns para trabajar todas las secuencias que obtuvimos haciendo búsquedas con HMMER. Este programa trabaja con secuencias que, de forma muy recomendable, ya deben de estar alineadas. Como la metodología trabaja analizando de forma preferencial aminoácidos hidrofóbicos, también se aconseja alinearlas con una matriz que evalúe las coincidencias encontradas cuando se alinean este tipo de aminoácidos (como la matriz de HCA).

Posteriormente, el programa procesa -como describimos en párrafos anteriores-, las similitudes en la distribución de aminoácidos hidrofóbicos contra el total de los mismos en una determinada secuencia (*HCA_Score*) [5, 8] y calcula la probabilidad de que las coincidencias no se deban a simple azar [5] (similar al valor de E de BLAST [13]).

Que todo el proceso se lleve a cabo requiriendo únicamente las secuencias, no sólo facilita la labor, favoreciendo con ello el análisis simultáneo de muchas secuencias; sino que además, permite eliminar el factor subjetivo durante la comparación de los diagramas de HCA, lo que aumenta el espectro de posibles aplicaciones para el uso generalizado de éste método.

HCA_Analyze_Multiple_Aligns produce como resultado cuatro archivos de salida por separado:

HCA_Analyze_Multiple_Align_Results:
Es el archivo que muestra los resultados de las comparaciones hechas entre las secuencias con base en las coincidencias entre los patrones de hidrofobicidad. Cada patrón que se encuentre (en un paso inicial del algoritmo) se compara con todas y cada una de las secuencias. Aquí se muestra los resultados de tales comparaciones separados en columnas que representan distintos parámetros. El conjunto de valores para los parámetros pueden ser interpretados como indicadores no-directos de homología entre las secuencias. El más importante dentro del algoritmo es el *HCA_Score*, que representa las diferencias/similitudes en cuanto al número y distribución de aminoácidos hidrofóbicos que existen cuando se comparan al menos dos secuencias.

Distinct_HCA_patterns:

Muestra en forma de lista los nombres de las secuencias que poseen un porcentaje de similitud de HCA menor al 60% (el *HCA_Score*) en comparación con todas las demás secuencias que tenemos en el alineamiento. Las secuencias enlistadas aquí presentan patrones estructurales únicos o distintivos. Este archivo de salida es exclusivo del programa que hace las comparaciones múltiples, HCAmA. Los resultados de esta lista corresponden a los Patrones Distintivos de Hidrofobicidad (PDH).

Minima:

Incluye y nombra a las secuencias (por el identificador que le corresponda en el archivo fasta) más divergentes en comparación con todas las demás dentro del alineamiento. A diferencia del listado en Distinct_HCA_patterns, aquí se resaltan las diferencias no sólo con base en el *HCA_Score*, sino que, además, se toman en cuenta los siguientes parámetros: Similitud de aminoácidos cargados y distribución de Prolinas. Este archivo de salida es exclusivo del programa que hace las comparaciones múltiples, HCAmA.

Memorize:

Finalmente, este archivo enlista el número total de secuencias analizadas, su longitud (máxima) y el número de patrones de hidrofobocidad distintivos encontrados.

Con la información contenida en estos archivos se puede evaluar la significancia estadística de las comparaciones entre secuencias en base a uno o varios parámetros. Esta comparación puede hacerse entre pares de secuencias o entre múltiples, y con ello inferir similitud estructural posiblemente debida a la homología entre secuencias.

5.7 Análisis de identidad entre los agrupamientos hidrofóbicos de las secuencias seleccionadas

Como se mencionó, el método de HCA es especialmente útil para trabajar con secuencias divergentes que no pueden ser analizadas por métodos tradicionales, sin embargo, una de las críticas más fuertes al mismo es que requiere de la intervención experta para interpretar y llevar a cabo las comparaciones entre los diagramas de HCA que se produzcan para las distintas secuencias de interés (recordemos que, para una secuencia que corresponde a una proteína, se obtiene un diagrama de HCA). Inicialmente, la comparación y/o análisis manual de los diagramas

era un componente fundamental e inevitable para el método, por lo que cuando las comparaciones requerían de llevarse a cabo entre un número elevado de secuencias/estructuras, había un enorme costo en tiempo y esfuerzo. Aunado a estas desventajas, el análisis de tales diagramas puede ser subjetivo, pues además de estar sujetas al criterio del investigador que analice los diagramas de HCA (en la determinación, de, por ejemplo, los límites de un agrupamiento hidrofóbico); con mucha frecuencia es complicado saber cuándo la similitud estructural es suficiente para concluir que es producto de ancestría común (homología), o que son coincidencias que se pueden esperar por puro azar (5, 6, 8). Naturalmente, este problema es especialmente notorio y difícil de superar cuando hacemos comparaciones entre secuencias divergentes que usualmente son filogenéticamente distantes.

Una forma de enfrentar estas limitaciones y a la vez fomentar el uso generalizado de un método que posee claras ventajas sobre aproximaciones más clásicas (ver sección “ventajas del método de HCA”); en el 2009, Pedro Silva propone un algoritmo que permite disminuir sustancialmente (y en una primera instancia eliminar por completo) la subjetividad del método, mientras que, simultáneamente, nos permite automatizar los análisis (5) así como evaluar la significancia de éstos. La premisa es que podamos obtener una medida conveniente de la similitud que existe entre patrones de distribución espacial de los aminoácidos hidrofóbicos (a nivel de secuencia, con el *HCA_Score*) considerando siempre la probabilidad de que pudiese existir una correlación con la distribución de los mismos aminoácidos en una secuencia al azar de composición/longitud similar (8). Así, la promiscua inserción de gaps para optimizar el alineamiento de secuencias divergentes podría favorecer identidades que no son producto del origen común entre las mismas (5); por lo que, tomando en cuenta la probabilidad de que, por simple coincidencia, al menos “n” aminoácidos hidrofóbicos se encuentren en un alineamiento de longitud “x” entre dos secuencias de tamaños/composición proporcionales, se convierte en un “factor de ajuste” con el cual podríamos definir la significancia de nuestros análisis. Por extrapolación, un algoritmo puntuador que evalúe los falsos positivos utilizando una aproximación sencilla durante el alineamiento, analizará la correlación que exista entre la longitud de “n” secuencias, la composición promedio de las mismas y el número de gaps que se introdujeron para alinearlas. Este es el mismo principio que utiliza el algoritmo HCA_Analyze_Multiple_Aligns para identificar los falsos positivos entre secuencias después de haberlas alineado. El acercamiento es especialmente robusto pues las comparaciones son pareadas entre las secuencias patrón (leer más adelante) y todas las demás que conforman el alineamiento completo, por lo que sólo aquellas genuinamente significativas serán exitosamente identificadas como tales.

Para el lector interesado, la formulación matemática detallada de los conceptos anteriormente expuestos se puede encontrar en el trabajo de Pedro Silva (Referencia 5).

Como ya podemos notar, es importante notar que el valor P de similitud para las secuencias debe ser tomado con cautela, pues es dependiente de la longitud del alineamiento (entre más grande sea, podríamos tener mayor número de coincidencias debidas simplemente al azar) y de la composición de las secuencias (5, 8, 65). Esto sucede así por la distinta tasa de sustitución que se presentan en las secuencias y la tendencia que tienen los programas que alinean a "optimizar" el acomodo global aun entre secuencias que no poseen origen común, lo que modifica el valor de la probabilidad total de que el alineamiento obtenido sea correcto (5, 6, 8). Como ya hemos mencionado, HCA_Analyze_Multiple_Aligns incluye una "función puntuadora" que establece un umbral para diferenciar las similitudes biológicamente significativas (homología) de los "artefactos" metodológicos introducidos por las secuencias en función de su tamaño, tasa de sustitución y nivel de divergencia [5]. Una característica distintiva de HCA_Analyze_Multiple_Aligns es que para hacer una discriminación más rigurosa de las secuencias similares de las que realmente no lo son podemos emplear el archivo de salida que produce el programa. Así podemos evaluar un número mayor de parámetros si éstos se consideran significativos para nuestro estudio, como podrían ser: la proporción de aminoácidos cargados, la distribución de Prolinas y el número total de aminoácidos hidrofóbicos. Llevar a cabo este análisis es completamente opcional y, aunque esto podría resultar complicado (aunque de igual forma se podría automatizar mediante programas (*scripts*) hechos por el mismo usuario) podría ser de gran utilidad, como se ejemplifica en el trabajo de Pedro Silva (ver Ref. 5), en donde se re-analizan casos en los que la similitud estructural es incorrectamente identificada como homología, lo que nos indica que el método de HCA, a pesar de sus bondades, produce falsos positivos que sólo pueden identificarse como tal hasta que se analizan una cantidad mayor de parámetros que pueden ayudar a discernir entre los casos más difíciles.

En párrafos anteriores, desglosamos el contenido de los archivos de salida que se obtienen con el programa HCA_Analyze_Multiple_Aligns (HCAmA). Dado que, por cuestiones de practicidad, en este trabajo no se realizarán las comparaciones pareadas de los 172 resultados positivos de un total de 735 secuencias; se decidió emplear un acercamiento práctico al problema utilizando los mismos resultados del programa. Con tal propósito, se utilizaron los resultados obtenidos y la lógica de las comparaciones que se encuentran detrás del archivo Distinct_HCA_patterns (que

enlista los Patrones Distintivos de Hidrofobicidad). Para hacerlo, obtuvimos un conjunto de secuencias homólogas distantes a la Lisozima Tipo C de Ave, las alineamos empleando la matriz de HCA, y, posteriormente, las analizamos con el algoritmo HCA_Analyze_Multiple_Aligns (HCAmA) tal cual fue descrito en la descripción del algoritmo. HCAmA nos encontró aquellos patrones de hidrofobicidad distintivos analizando todas y cada una de las secuencias presentes en nuestro alineamiento, al mismo tiempo que nos permitió identificar aquellas secuencias consideradas positivas, esto es, que la comparación de "x" patrón contra "y" secuencia fuese significativa en términos de la composición/distribución de los aminoácidos más hidrofóbicos (VILFMYW).

Aplicar el algoritmo HCAmA nos permite hacer las comparaciones a gran escala, por lo que con las 172 secuencias se construyó un dendograma estructural utilizando un algoritmo de distancia, Neighbor-Joining. Posteriormente, los 49 Patrones Distintivos de Hidrofobicidad (PDH) identificados por el algoritmo HCAmA fueron mapeados en las ramas del dendograma. Con ello pretendíamos ser capaces de observar gráficamente la distribución de los patrones PDH en un dendograma estructural [Figura 7].

5.8 Los Patrones Distintivos de Hidrofobicidad (PDH)

La búsqueda de los PDH que se reportan en la lista Distinct_HCA_patterns correspondientes a secuencias únicas en el alineamiento, son uno de los pasos iniciales del algoritmo (Figura 7); por lo que, de hecho, HCA_Analyze_Multiple_Align_Results es el archivo que muestra los resultados de las comparaciones hechas entre todas las secuencias presentes en el alineamiento contra todos y cada uno de los Patrones Distintivos de Hidrofobicidad (PDH) que se encontraron dentro del mismo.

Estos patrones representan secuencias únicas que nos podrían estar reflejando la diversidad estructural de nuestras proteínas tomando en consideración los parámetros más importantes dentro del método de HCA (número de aminoácidos hidrofóbicos totales y/o compartidos, distribución espacial de los agrupamientos de aminoácidos hidrofóbicos, distribución de aminoácidos cargados, distribución de Prolinas, etc.). Para construir el listado de PDH, el algoritmo HCAmA inicia realizando una comparación en parejas (*pairwise*). Selecciona la primer secuencia del alineamiento y busca otra (dentro del mismo alineamiento) que tenga un *HCA_Score* menor al 60%; esto es, que sea un 60% más diferente (en cuanto a la distribución y composición de los aminoácidos hidrofóbicos: VILFMYW) a ésta primer secuencia (Figura 7).

Luego, el algoritmo busca de nuevo una tercer secuencia que tenga un *HCA_Score* menor al 60% que cualquiera de las dos secuencias seleccionadas previamente. Este último paso se repite hasta que todas las secuencias del alineamiento hayan sido analizadas (Figura 7).

Al final, en el archivo se nombran las secuencias que forman "Patrones Distintivos de HCA" (que corresponden a los PDH), pues, como lo hemos mencionado, el número y distribución de los aminoácidos hidrofóbicos son significativamente distintos a los que presentan las demás secuencias en el alineamiento.

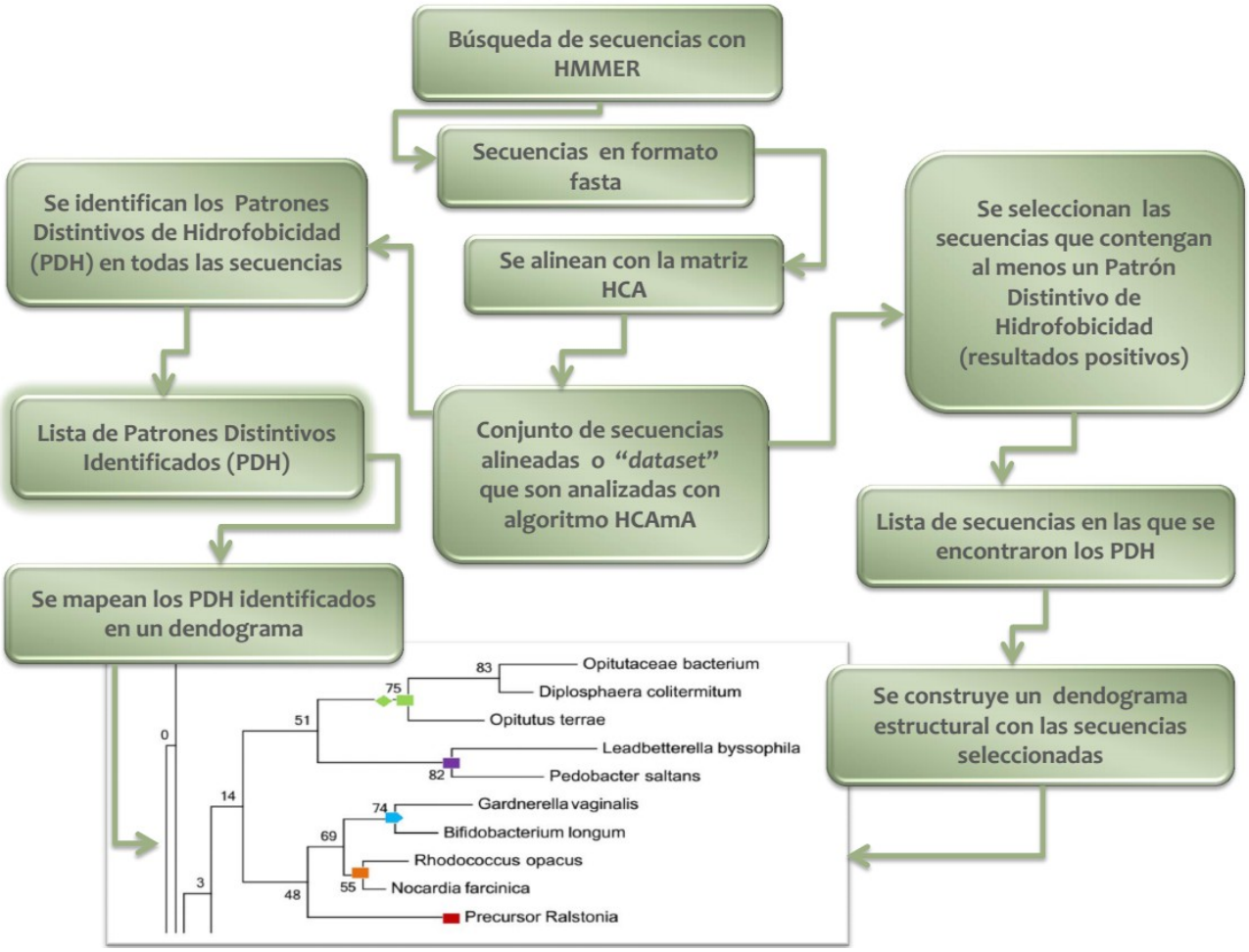

Figura 7. Esquema de la metodología seguida. Se comienza con la búsqueda de homólogos distantes de la Lisozima Tipo C. A las secuencias resultantes se les aplica un filtro de redundancia (por ejemplo, se eliminan las secuencias 99% idénticas y las múltiples copias que un Género de Bacterias pueda tener para la secuencia templado). Posteriormente, se alinean con la Matriz de HCA. Este alineamiento es el conjunto de datos que se analiza mediante el programa HCAmA. Como resultado, se obtienen dos archivos. El primero, contiene valores numéricos que reflejan la calidad de las comparaciones a nivel de secuencia de los agrupamientos hidrofóbicos. El segundo, es un listado que corresponde a los Patrones Distintivos de HCA. Los patrones reportados en esta lista, se mapean en un dendograma construido con las 172 secuencias comparativamente significativas obtenidas en el primer archivo. La visualización de los patrones en el dendograma, nos permiten hacer inferencias de similitud a nivel estructural, a la vez que nos permite resumir una gran cantidad de información de manera eficiente y concreta.

En la literatura (5, 8, 58, 64) se reporta que hay evidencia para suponer que la estructura de una gran cantidad de proteínas es similar en cuanto al número de agrupamientos de aminoácidos hidrofóbicos que les componen; lo que a su vez, nos permite suponer que muy posiblemente aquellas más parecidas en este aspecto también podrían tener funciones y estructura tridimensional similar; por lo menos en cuanto a la proporción y tipo de estructuras secundarias que poseen (6, 8, 38, 50).

6. RESULTADOS

Obtuvimos 10546 secuencias con HMMER después de cinco iteraciones. En ésas 10546 secuencias había 9742 de Bacterias, 475 de Eucariontes, 59 secuencias provenientes de Virus (fagos) y 270 secuencias sin clasificar. Aunque se utilizó la base de datos no-redundante del NCBI para la búsqueda de secuencias, en lo resultados se obtuvieron muchas secuencias repetidas, así que para reducir el número de comparaciones redundantes y analizar sólo secuencias que entre ellas fueran significativamente distintas, se hizo una limpieza inicial que consistió en dejar una secuencia correspondiente a sólo una especie de cada género en la que se haya encontrado el gen de la Lisozima Tipo C (p. ej. sólo se mantuvo a *Helicobacter pylori* como única representante del género. Todas las demás 371 especies o cepas del género *Helicobacter* no fueron seleccionadas para incluirse en el análisis). La selección de qué secuencia/especie se quedaría en el conjunto de datos a analizar (*dataset* final) se hizo con base en el valor de "E" y a la puntuación de significancia (*bit score*), según los valores numéricos obtenidos de HMMER. Al final de este procedimiento, el número total de secuencias fue reducido a 729.

Estas 729 secuencias ya curadas (libres de secuencias redundantes) fueron analizadas con el algoritmo HCA_Analyze_Multiple_Aligns (HCAmA), con lo que obtuvimos como resultado 229 secuencias por encima del umbral de aceptación (resultados positivos) y un total de 49 patrones distintivos. Eliminando aquellas comparaciones positivas que se consideraban como significativas entre sí mismas [referirse a las discusiones para una explicación elaborada de este artilugio metodológico del algoritmo] nos quedamos con 172 secuencias positivas verdaderas.

Subsecuente a estas múltiples depuraciones de secuencias repetidas, mapeamos los 49 Patrones Distintivos de Hidrofobicidad (PDH) obtenidos con el algoritmo HCAmA sobre un dendograma que se construyó con las 172 secuencias finales positivas. Quisiéramos resaltar que el dendograma se

construyó manteniendo los *gaps* que se insertaron durante el alineamiento original de las 729 secuencias. Estos *gaps* previamente insertados no fueron “reseteados” para hacer un nuevo alineamiento. Esto se decidió así porque nuestro interés no era construir una “filogenia”, sino obtener un dendograma hecho con un algoritmo que nos permitiera analizar la similitud estructural entre agrupamientos de HCA y con ello delimitar cuidadosamente las agrupaciones y patrones que éste dendograma nos mostraba (usamos Neighbor Joining [70]), pues puede analizar la distancia genética, -concepto que es fácilmente extrapolable como un indicador de similitud entre secuencias- [70, 71]).

Eliminar los *gaps* originales (dentro del alineamiento de las 729 secuencias) y comenzar el análisis de nuevo (con lo que se insertarían “nuevos” *gaps* de acuerdo a la nueva muestra) para mejorar el óptimo global del alineamiento de sólo las 172 secuencias aceptadas seguramente generaría agrupaciones nuevas que ya no serían descriptivas de la diversidad estructural que estaba contenido en el alineamiento original de las 729 secuencias.

Figura 8. Dendograma estructural. Podemos notar que el dendograma consta de: 90 secuencias de *Proteobacterias*, 29 secuencias de *Bacteroidetes*, 16 secuencias de *Actinobacteria*, 19 secuencias de *Firmicutes* y 18 secuencias de algunos otros órdenes representados en menor proporción, como, *Verrucomicrobia* (5), *Synergistetes* (5) y *Deinococcus* (4), otros órdenes, como *Fusobacteria* y *Planctomycetes* sólo tienen una secuencia representante. El árbol también posee una única secuencia del Fago de *Escherichia coli*. **El árbol completo en un tamaño legible puede apreciarse con mayor detalle en la sección: Anexos, de este trabajo.**

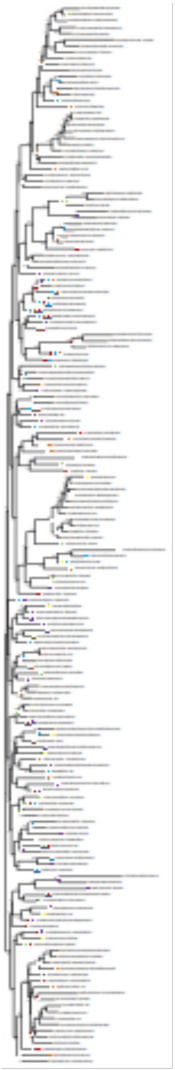

Tabla 3. Listado de los 49 Patrones Distintivos de Hidrofobicidad (PDH). En estas columnas, se muestran el símbolo y color de cada uno de los 49 PDH, así como su nombre y representatividad. La representatividad de los PDH (tercera columna) hace referencia al número total de secuencias en las que se identificaron dos o más PDH. Los patrones que sólo se encuentran en una secuencia portadora son las secuencias *"Singlets"*.

Símbolo	Nombre de Patrón	No. Secuencia Portadora
	Cardiobacterium	2
	Anaplasma	1
	Myxococcus	2
	Sulfitobacter	1
	Anaeromyxobacter	14
	Asticcacaulis	1
	Leptothrix	5
	Acidobacterium	19
	Methylobacillus	5
	Holophaga	2
	Methylophaga	6
	Geobacillus	11
	Mariprofundus	15
	Pyramidobacter	2
	Thiothrix	1
	Fusobacterium	1
	Thermovibrio	3
	Thiocystis	5
	Thiorhodovibrio	6
	Meiothermus	9
	Austwickia	11
	Rhodobacteraceae	13
	Beggiatoa	1
	Desulfobacula	3
	Nitrospina	1

Continuación...

Símbolo	Nombre de Patrón	No. Secuencia Portadora
	Thermincola	1
	Desulfomicrobium	1
	Tannerella	24
	Paraprevotella	3
	Caldithrix	1
	Erythrobacter	4
	Ketogulonicigenium	17
	Sphingopyxis	4
	Methylocystis	3
	Thermus	1
	Ralstonia	1
	Diplosphaera	3
	Opitutus	3
	Leadbetterella	2
	Rhodococcus	2
	Bifidobacterium	2
	Plesiocystis	1
	Phycisphaera	1
	Desulfarculus	1
	Conexibacter	2
	Rhodovulum	4
	Rhodomicrobium	3
	Parvularcula	3
	Pedosphaera	2

7. DISCUSION

7.1 Las Lisozimas de Bacterias como modelo de análisis estructurales

Desde los años 80 ha existido la duda respecto al origen y parentesco de las 100 diferentes familias de Glucósido-hidrolasas, súper familia a la que pertenecen las enzimas Lisozimas (23, 66). Estudios específicos (16, 50, 53, 66, 67, 73) se han llevado a cabo para tratar de encontrar similitud que se pueda interpretar como homología, y, aunque los resultados son variables, el consenso

indica que, al menos a nivel estructural, la mayoría de las Lisozimas estudiadas poseen características que ciertamente apuntan a un muy plausible origen común (16, 17, 66, 73). Anteriormente, las limitaciones en cuanto a las secuencias disponibles o a las estructuras tridimensionales ya resueltas impedía realizar un gran número de comparaciones, por lo que era difícil hacer inferencias de gran poder predictivo, así que por mucho tiempo el posible origen común de toda esta gran diversidad de Lisozimas sólo se mantuvo como una hipótesis basada en evidencia limitada (16, 49, 53, 67).

En la actualidad, el panorama es distinto. Ahora contamos con una enorme cantidad de información, el problema es que la misma dificulta el análisis detallado de las distintas Lisozimas, y en definitiva, el costo y dificultad que podría representar resolver las estructuras de una muestra que resulte significativa para hacer inferencias es prohibitivo. La opción que parece más viable es explotar al máximo la información con la que contamos. En las bases de datos se encuentran muchas secuencias pero con pocas estructuras resueltas, así que las dificultades que se pueden encontrar para llevar acabo análisis conjuntos (de secuencia y estructura) ciertamente podría estar limitando nuestro conocimiento respecto a múltiples aspectos (incluyendo la evolución) de éstas enzimas. El reto, ahora, es entender cómo podemos sacar el máximo provecho de tal información.

Aquí presentamos como una opción viable el método de Pedro Silva (5) aplicado a un problema que lleva mucho tiempo sin encontrar una respuesta confiable; este es, la diversificación y variabilidad de las Glucósido-hidrolasas de la familia 22, las enzimas Lisozimas. Estas enzimas no representan un problema trivial, lo que nos lleva a resaltar las siguientes observaciones respecto a éstas enzimas:

1) La actividad catalítica es llevada a cabo por sólo dos aminoácidos, (algunos proponen un tercero, [53]) y, aunque usualmente invariantes, pueden encontrarse a distancias muy distintas en la secuencia, e inclusive, cambiar de posición a nivel estructural en diversas Lisozimas (53, 66, 67). Por esta razón, los algoritmos -aun con matrices dinámicas- pueden fallar en la labor de encontrar óptimos locales durante el alineamiento, lo que se traduce en dificultad para encontrar dominios o motivos conservados entre comparaciones de múltiples Lisozimas. Que la mayoría de las Lisozimas sean multidominio complica aún más esta labor, al mismo tiempo que nos indica que un estudio apropiado de las mismas implica la capacidad para lidiar con tal modularidad (algo que muchas metodologías actuales no pueden enfrentar).

2) Aun si consideramos que se ha logrado ubicar el dominio glucósido-hidrolasa (que define a la mayoría de las Lisozimas estudiadas, pero también a otras enzimas que no pertenecen a la misma familia (76); el motivo/dominio que es el encargado de "ubicar" en la posición correcta a los aminoácidos que llevan a cabo la catálisis es altamente variable (66). En ocasiones, esta variabilidad está en función de la estructura tridimensional del sustrato, por lo que enzimas que actúan sobre diversos sustratos -como las Lisozimas-, oscurecen los resultados de ensayos bioquímicos que ayudan en su identificación. Naturalmente, esto imposibilita la delimitación de las Lisozimas de acuerdo a su función.

3) A nivel de secuencia, algunas Lisozimas son tan variables que difícilmente se pueden identificar como homólogas a menos que se utilice algún tipo de información estructural (53, 54, 66, 67, 75). En ocasiones, esta variabilidad es capaz de producir similitud "engañosa". Cuando esto sucede, es posible que en los pasos iniciales de un alineamiento se agrupen secuencias que pertenecen a organismos filogenéticamente distantes. En este punto, quisiéramos enfatizar que durante la realización de este trabajo, pudimos notar discordancias filogenéticas notables, como, por ejemplo, que una Lisozima Tipo C de Eucarionte tenga un porcentaje de identidad mayor contra una de Fago (al menos parcialmente) que con otra de algún otro Eucarionte, como la Lisozima Tipo C de *Pavo cristatus* (Pavo real). Creemos que la variabilidad a nivel de secuencias y tasa acelerada de mutación a la que están sujetas algunas de las secuencias (principalmente parálogas) de las Lisozimas, podrían estar produciendo estos artilugios metodológicos.

4) Las Lisozimas presentan regiones de similitud estructural local muy marcada, pero la misma es difícil de detectar entre las representantes más divergentes, aun cuando éstas pertenezcan a organismos filogenéticamente emparentados (66, 67).

5) Se ha encontrado que en el genoma de muchos mamíferos hay abundantes copias de genes que codifican para algún tipo de Lisozima (16, 17). La propuesta es que la tasa de sustitución de muchas de estas secuencias, así como su ubicación aparentemente "desordenada" dentro del genoma de algún organismo, se puede tomar como un indicador de que, en realidad, se trata de genes xenólogos. Por el momento, la historia evolutiva de dichas enzimas (de las Lisozimas "ancestrales") es desconocida (53, 54, 75), pues no ha sido factible identificar correctamente los genes ortólogos para llevar a cabo las reconstrucciones filogenéticas que muy posiblemente arrojarían información valiosa para resolver la cuestión (16).

Los intentos que se han hecho, sólo toman en cuenta secuencias con más del 40% de identidad, lo que, de nuevo, excluye del análisis a las proteínas más divergentes (16).

6) En este trabajo hipotetizamos que, aun si contáramos con los genes ortólogos de las Lisozimas en estudio, tendríamos el gran problema de que la filogenia resultante (una genealogía) muy posiblemente mostraría cierto nivel de discordancia con una filogenia de especies (algo que, de hecho, es evidente con nuestro dendograma estructural, [Figura 8]), fenómeno que es muy observado cuando el marcador molecular tiene una tasa de sustitución muy elevada, sufre duplicaciones frecuentes, las secuencias son transferidas horizontalmente y/o la topología del filograma resultante muestra evidencia de que las secuencias han sufrido un fenómeno conocido como reemplazo de genes no-ortólogos, lo cual es perfectamente compatible con lo que esperamos podría ocurrir a las secuencias de Lisozimas (16, 17, 38, 73).

En el caso de secuencias con un porcentaje de identidad muy aceptable (mayor al 85%), la identificación de aminoácidos esenciales para la actividad catalítica, así como para mantener una estructura tridimensional estable, se puede lograr con un alineamiento bien cuidado, lo que implica una correcta elección de la matriz que evaluará las sustituciones, así como penalizaciones a la apertura y extensión de *gaps* acorde con las secuencias que se evaluarán (53, 17). Sin embargo, con las secuencias más divergentes, los patrones de conservación de aminoácidos, inclusive los patrones de variabilidad, sólo pueden ser evaluados objetivamente a niveles estructurales (49, 75).

7.2 Problemática con las Lisozimas de Bacterias.

Como ya lo mencionamos en resultados, haciendo búsquedas con JackHmmer (22), obtuvimos 10546 secuencias de Lisozimas. Los parámetros de búsqueda enfatizaban la identificación de secuencias homólogas distantes. Interesantemente, JackHmmer encontró 9742 secuencias que pertenecían a bacterias, algunas de ellas con valores de expectativa (*E-value*) altos y de *BitScore* bajos, que enfatizaban las diferencias a nivel de secuencia.

La abundancia de estas secuencias -aun en las últimas rondas de la búsqueda iterada-, nos hizo cuestionar: ¿por qué existen tan pocos trabajos en los que se estudia a las Lisozimas de bacterias a pesar de la evidente disponibilidad de tales secuencias?. Pensamos que la respuesta radica en que

la mayoría de los estudios son sólo de tipo estructurales, y las súper-imposiciones suelen llevarse a cabo sólo entre las Lisozimas más estudiadas (no olvidemos que, generalmente, son las únicas que cuentan con estructura resuelta).

En muchos de estos estudios la constante es la declaración de que, a nivel de secuencia, prácticamente no existe semejanza, mientras que a nivel estructural, hay características, que si bien similares, es posible que se deban sólo al azar y por ello no es posible considerarlas certeramente como producto de un origen común (66, 67). Hacer inferencias y estudios con enzimas muy divergentes (en donde algunas son más y otras menos) resulta complicado, pues el valor predictivo de tales estudios se ve altamente comprometido. Los análisis más tradicionales, como la inferencia filogenética, suelen llevarse a cabo utilizando solamente secuencias que presentan, al menos 40% de similitud entre ellas (16), por lo que los estudios suelen arrojar resultados con buen sustento pero muy sesgados (38), ya que, deliberadamente, se omiten aquellas secuencias variables que suelen resultar problemáticas, pero que definitivamente podrían resultar no sólo complementarias sino altamente informativas (16).

El acercamiento que este algoritmo nos permite es único y muy robusto, pues, aunque se trabaja sobre la secuencia, el análisis es estructural. Recordemos que desde el paso inicial en el alineamiento, las secuencias fueron alineadas con una matriz especial (Matriz de HCA) que alinea en base a las coincidencias entre agrupaciones (*clusters*) de aminoácidos hidrofóbicos. Con éstos aminoácidos hidrofóbicos podemos hacer inferencias estructurales en base a la presencia, cantidad y ubicación de los agrupamientos que forman, pues ya se ha confirmado que éstos son de gran utilidad para detectar homología entre secuencias filogenéticamente distantes (4, 6, 8, 42, 50, 58, 64). Fue así que el algoritmo nos permitió enfocarnos en las Lisozimas de bacterias, que son las Lisozimas menos estudiadas y para las cuales la información estructural es sumamente limitada.

El estudio detallado de cada una de las Lisozimas bacterianas incluidas en el trabajo está fuera de los alcances pretendidos dentro de nuestros objetivos; sin embargo, pudimos confirmar de forma gráfica lo que se ha reportado en literatura (23, 24). Son secuencias altamente variables y de muy diversos tamaños. Los alineamientos tradicionales (hechos sin la matriz de HCA) requieren una enorme cantidad de *gaps* y ello sólo para favorecer las coincidencias entre regiones pequeñas (no más de 10 aminoácidos, en muchos casos). En la sección “Resultados” de este trabajo ya hemos mencionado que una gran cantidad de secuencias aparecen como sin clasificar, inclusive, en algunos casos la misma es dudosa (23).

Con la intención de hacer un acercamiento inicial en tal aspecto, hicimos búsquedas con JackHmmer para las primeras 5 de las 270 secuencias que se encontraban como “sin clasificar” en la base de datos no-redundante del NCBI (resultados no mostrados). En todos los casos, algunas de las mejores coincidencias (*hits*) para tales secuencias corresponden a representantes bacterianos. Debido al número tan reducido de las secuencias comparadas y a la aproximación tan sesgada que hicimos del fenómeno, realmente no es posible generalizar sobre los resultados, por lo cual es muy difícil saber si ésas 270 secuencias pertenecen a bacterias cuyo genoma no ha sido secuenciado completamente o no están correctamente anotadas. Cabe resaltar, que el algoritmo de HCA no encuentra patrones distintivos en el alineamiento inicial que incluía a dichas secuencias “sin clasificar”; esto, aunado a que el algoritmo de HMMER encuentra el dominio transglicosilasa característico de una gran cantidad de Lisozimas (76) podemos tomarlo como evidencia, al menos indirectamente, de que éstas secuencias son, efectivamente, Lisozimas bacterianas. Enzimas con marcadas características de Lisozima (tan marcadas que incluso HMMER puede detectarlas) no estén correctamente clasificadas o por lo menos identificadas como posibles glucósido-hidrolasas, pensamos que puede interpretarse como evidencia a favor de que las Lisozimas de bacterias han sido poco estudiadas, y esto se refleja en que existe limitada información respecto a su clasificación (16, 17, 23).

En las Lisozimas de virus y bacterias más estudiadas (muy pocas), el sitio catalítico es variable; en ocasiones faltan los aminoácidos responsables (17, 75) o están en diferentes posiciones, ya sea en el alineamiento o en la estructura tridimensional (53, 54, 66, 67, 75). Los dominios tampoco están del todo conservados; algunas Lisozimas tienen estructuras en las que predominan las alfa/beta, otras son predominantemente alfa-hélices interconectadas mediante ”asas” o *loops* hipervariables con hojas beta, y en ocasiones, la estructura central es un TIM-Barrel que presenta asociaciones con diversas estructuras secundarias de conexión (*coiled-coils*) y/o plegamientos característicos, como el TIM-Barrel “central” que poseen en su estructura una gran cantidad de Lisozimas (77). Por si fuera poco, las Lisozimas de bacterias y fagos presentan el fenómeno de modularidad (54, 75), que hace referencia a la hipótesis de que muchas Lisozimas están constituidas por repeticiones del mismo o diferentes dominios; inclusive, algunos que no son comunes dentro de la familia de las Lisozimas, lo que dificulta la detección e identificación de motivos/dominios conservados.

7.3 Utilidad de los Patrones Distintivos de Hidrofobicidad (PDH)

La cuestión es, cómo enfrentamos el estudio de enzimas tan variables y divergentes. Es evidente que la identificación de características muy comunes no es la opción, pues de esta manera sólo podemos estudiar secuencias/estructuras de proteínas muy similares, lo que excluye a aquellas que, a pesar de tener un origen común, han divergido tanto que ya no poseen las características ancestrales más conservadas. Analizar tomando en cuenta toda la variabilidad existente tampoco parece ser lo más acertado -además de ser altamente impráctico-, pues si tenemos una enorme diversidad y por ello estamos dispuestos a trabajar con todas las secuencias/estructuras más variables, se vuelve muy complicado delimitar hasta qué punto ésa variabilidad es todavía producto de la divergencia y evolución de las secuencias/estructuras a través del tiempo. En este caso, no podemos ignorar el hecho de que esta flexibilidad "juegue" en contra y facilite la inclusión de secuencias/estructuras que en realidad no poseen un origen común, lo que por supuesto, podría tener implicaciones no previstas para el estudio.

Aquí, consideramos que la respuesta podría recaer en analizar la predisposición de éstas proteínas a formar determinada estructura tridimensional. Ya no evaluaríamos una presencia/ausencia de caracteres estructurales, sino la capacidad de producirlas, capacidad que, esperamos, sea aquella característica intrínseca heredable, producto de la ancestría/descendencia. Si estamos en lo correcto, la presencia de las estructuras ya no define el posible origen común de las secuencias/estructuras; es, más bien, el resultado de la influencia de la selección natural y las presiones diferenciales a la que están sujetas a través del transcurso del tiempo. Para analizar esta capacidad de formar estructuras que adopten un plegamiento característico, similar y posiblemente con funciones análogas, el algoritmo se enfoca en la distribución de los aminoácidos hidrofóbicos.

El análisis de los patrones distintivos de hidrofobicidad puede ser altamente informativo, pues, aun concentrándonos en el análisis de sólo los aminoácidos hidrofóbicos, podemos extrapolar los resultados y obtener información sobre la tendencia que tienen las proteínas a formar estructuras irregulares, pues aquellas regiones carentes de aminoácidos hidrofóbicos que no estén correlacionadas con la formación de estructuras secundarias podrían estar participando en la formación de loops o estructuras de interconexión entre dominios, las cuales, poseen una asociación estadística muy sólida con los aminoácidos hidrofílicos, como lo son P, G, D, N, S (42). Por analogía, si una estructura es similar en cuanto a la distribución espacial de sus aminoácidos

hidrofóbicos, también es posible que lo sea en las regiones “desordenadas” usualmente abundantes en aminoácidos polares o hidrofílicos, pues la hidropatía (tendencia a repeler o aceptar la interacción con moléculas de agua [76]) es una característica mutuamente excluyente; es decir, una misma región no puede ser hidrofóbica e hidrofílica en el mismo instante; sin embargo, esta dicotomía entre aminoácidos polares y no-polares, es la que se piensa, es primordial para el plegamiento tridimensional de las proteínas (4, 38, 42, 76). Dentro del contexto que veníamos manejando en párrafos anteriores, esto favorece un tratamiento más realista de la diversidad en el que ya no se analiza con base en presencia o ausencia de caracteres estructurales, sino a la capacidad de formar tales estructuras (la característica heredada, que nos habla de un posible origen común).

Entonces, aplicado a 729 secuencias en las que solamente había una especie representante de un único género -esperando que con esto pudiésemos reducir la redundancia del estudio-, obtuvimos 49 patrones distintivos de HCA, pero sólo 172 resultados positivos o por encima del umbral de aceptación; lo que esto quiere decir es que son secuencias significativas para llevar a cabo comparaciones múltiples. Esta disminución de los resultados corresponde a que el algoritmo considera como resultados significativos las comparaciones pareadas entre las secuencias patrón y la secuencia en sí misma (en la literatura se suelen denominar “*Singlets*”). Como son las secuencias las que poseen los patrones distintivos, cuando un patrón es comparado contra la secuencia que le dio origen, el resultado es positivo, pues existe concordancia en cuanto a la distribución de sus agrupamientos de aminoácidos hidrofóbicos; sin embargo, esto es un artilugio del algoritmo que puede ser ignorado (5), pues comparar una secuencia contra ella misma y encontrar algún patrón (de cualquier tipo) es completamente carente de utilidad.

Ahora, sería posible pensar que sólo unos cuantos patrones distintivos de HCA son compatibles con una gran cantidad de “tendencias” hacia la adquisición de determinados plegamientos (predisposición que en la literatura ha comenzado a denominarse como “el foldoma”) de una gran cantidad de representantes de diversas familias de proteínas globulares (clasificadas en base a la estructura tridimensional), cuestión que posiblemente nos esté reflejando que cualidades o características no tan variables, como lo es la distribución de los aminoácidos hidrofóbicos (5, 50), sean los responsables de limitar (al menos en parte) el universo de posibles plegamientos, aun de las proteínas globulares más distintas!!? (38, 50, 65, 73).

En apoyo a los resultados que estamos obteniendo, Orengo *et al*. en 1994 proponen que existen lo que ellos llaman súper-plegamientos ("*superfolds*"); plegamientos tan comunes que son capaces de describir el 30% de todas las proteínas globulares (77). Uno de ésos súper-plegamientos, el TIM-Barrel, es una estructura que predomina en algunas Glucósido-hidrolasas, pero la divergencia y eventual degeneración hacia la adopción de estructuras interconectadas por distintas estructuras secundarias es lo que parece preponderar en las Lisozimas, por lo que, consideramos nosotros, analizar la composición y distribución de, principalmente, los siete aminoácidos más hidrofóbicos (VILFMYW), que en conjunto podrían influenciar la capacidad de adoptar un plegamiento determinado, podría ser más informativo que identificar la presencia/ausencia de estructuras/dominios tridimensionales, en especial si el enfoque es el estudio de secuencias/estructuras altamente versátiles.

7.4 Detalles importantes de las secuencias analizadas

Como dato interesante, a diferencia de PSI-BLAST, en HMMER, los resultados que se obtienen en las últimas iteraciones suelen ser más significativos que los primeros (22). Considerando que empleamos la Lisozima Tipo C de un ave como secuencia templado (*Query*), hipotetizábamos que las secuencias homólogas más distantes que el algoritmo encontraría corresponderían a Bacterias y/o Fagos, -algo que sí ocurrió-, por ello, nuestro dendograma posee 171 secuencias de Bacterias y una de Fago (el de *Escherichia coli*); sin embargo, también esperábamos que los resultados más significativos (en términos de los estadísticos de confianza del algoritmo de HHMER) fueran secuencias de Lisozimas pertenecientes únicamente a Eucariontes. Esto último no se cumplió; de hecho, nuestro dendograma construido casi en su totalidad con Lisozimas de bacteria, incluye alrededor de un 10% de las secuencias encontradas en la última ronda de iteración en HMMER. No seleccionamos todas las secuencias que encontramos en la última iteración por dos razones principales:

1) Nosotros tenemos como objetivo analizar secuencias homólogas distantes de Lisozima; por lo mismo, consideramos que las Lisozimas de bacteria encontradas en la última iteración tendrían gran posibilidad de ser homólogas cercanas de nuestra enzima templado (con base en, nuevamente el valor de "E" y *Bit-Score*), por lo que su utilidad para cumplir nuestros objetivos podría ser limitada.

2) Muchas de estas secuencias resultaron ser secuencias repetidas de aquellas con un valor de “E” muy bajo y un *BitScore* muy alto, y recordemos que para evitar redundancias en los resultados, se hizo una depuración inicial de secuencias; por tal motivo, muchas de las secuencias repetidas (aun aquellas con buenos valores estadísticos) fueron descartadas en el estudio.

Nos resulta muy interesante que el algoritmo (HMMER en especial, que es más capaz de buscar secuencias distantes y/o muy divergentes) nos indique que la secuencia de la Lisozima Tipo C de un ave (una Lisozima Eucarionte) sea estadísticamente similar a la secuencia de las Lisozimas de Bacterias. En la literatura está altamente debatido si todas las Lisozimas son homólogas a pesar de su divergencia a nivel de secuencia, y todavía más, cuál, de las diferentes versiones que parecen existir, es la que podemos considerar como la versión ancestral; o por lo menos, más similar a la versión ancestral. Así que, ¿Es posible que HMMER nos esté mostrando evidencia de que las Lisozimas de Eucariontes y Bacterias tienen un origen común basados en la similitud a nivel de secuencia?. Recordemos que la similitud estructural es lo que define a las Lisozimas, así como la evidencia mejor estudiada que apoya con mayor fuera la hipótesis de homología; algo que no sucede con la secuencia.
Aunque los resultados de la búsqueda con un algoritmo de similitud (como lo son HMMER o BLAST) son insuficientes para invocar un posible origen común y con ello resolver cuestiones de homología, este hecho, que si bien podría no ser más que un artilugio metodológico, nos llamó la atención lo suficiente como para hacerlo notar.

7.5 Agrupamientos estructurales

Como ya lo hemos enfatizado en párrafos anteriores, el nuestro, es un dendograma estructural que muestra la relación que existe entre las secuencias, los aminoácidos hidrofóbicos que poseen y las estructuras secundarias que pueden formar. Lo que la topología del dendograma muestra, son las agrupaciones entre las secuencias de acuerdo a la similitud que poseen en términos de la ubicación, composición y longitud de sus agrupamientos de aminoácidos hidrofóbicos. Lo anterior no implica que aquellas secuencias que formen un agrupamiento (o un “clado”) posean el mismo PDH o que necesariamente tiendan a adoptar una estructura tridimensional idéntica; más bien, lo que este agrupamiento de ramas indica, es que existe similitud espacial entre sus agrupamientos de aminoácidos hidrofóbicos, por lo que podrían estar formando estructuras secundarias regulares similares (Figura 8). Estas estructuras secundarias regulares, en conjunto, pueden servir de “mapa

topográfico" para predecir el plegamiento que adoptará determinada proteína, pero de ninguna manera indican que poseen la misma estructura tridimensional.

En trabajos en los que se usa un dendograma como guía para establecer similitud estructural entre proteínas, se ha observado que suele existir cierto nivel de concordancia filogenética con una filogenia de especies (16, 17). Esto significa que los patrones estructurales de las proteínas de diversos organismos suelen ser correspondientes con la clasificación taxonómica de los organismos o especies portadoras. En nuestro caso, observamos una tendencia similar.

El dendograma muestra regiones (o sub-árboles) en los que se agrupa casi exclusivamente a Proteobacterias, mientras que otras agrupan sólo a *Bacteroidetes*, etc. En las regiones en las que se agrupan a organismos de distintos órdenes observamos una gran prevalencia de secuencias con ramas largas. En este trabajo no discutiremos a fondo (pues está muy fuera de nuestros objetivos) el fenómeno de atracción de ramas largas y las implicaciones que pueda tener en las agrupaciones mostradas en el dendograma; sin embargo, creemos que las similitudes azarosas que puede haber en cuanto a la composición de aminoácidos hidrofóbicos entre este tipo de secuencias, ciertamente podría estar favoreciendo la aglomeración de secuencias que pertenecen a distintos órdenes. Notablemente, aunque la técnica del *bootstrap* (fuertemente criticada por ser inconclusa bajo escenarios difíciles (82), en general, muestra valores de soporte bajos en las ramas del dendograma (en más del 70% del árbol son menores al 50%, Figura 8), resulta de utilidad para identificar aquellas regiones que poseen concordancia filogenética dentro del dendograma, esto es, el dendograma estructural muestra ramificaciones similares a las que se observan en una filogenia de especies.

Recordemos que las agrupaciones (clados) muestran las similitudes que existen en la composición de aminoácidos hidrofóbicos entre la secuencias involucradas, por lo que la interpretación del *bootstrap* en tales ramas, no se limita a ser sólo una medida de soporte de las mismas; así que, sumado a la gran variabilidad de las secuencias con las que el dendograma fue construido, consideramos que los valores bajos pueden ser resultado de artilugios metodológicos, como: **a)** Alineamiento sesgado, **b)** No todos los sitios del alineamiento son informativos, **c)** Como el *bootstrap* es un muestreo con reemplazo, se pueden estar perdiendo sitios de importancia vital para el soporte de determinadas ramas, **d)** Es sabido que, en general, para los análisis de secuencias altamente divergentes mediante distancias genéticas (NJ), el soporte de ***bootstrap*** es bajo y puede subestimar la "veracidad" de los clados [Brocchieri, 2001].

7.6 Representatividad de los PDH

Los PDH [Patrones Distintivos de Hidrofobicidad] son arreglos “únicos” de acuerdo a la distribución y abundancia de aminoácidos hidrofóbicos que podemos encontrar en las secuencias de las proteínas. Dado que existe evidencia de que los aminoácidos hidrofóbicos juegan un papel preponderante en la estructura de las proteínas, las secuencias de una enorme cantidad de proteínas pueden ser evaluadas en términos de la composición de tales aminoácidos (8, 38, 50, 65). Gracias a lo anterior, la utilidad más notoria de estos patrones de hidrofobicidad -que encontramos haciendo uso del algoritmo de HCAmA- es que nos permiten identificar características comunes entre secuencias que resultan filogenéticamente distantes, por lo que podrían ser una herramienta confiable para afrontar el estudio estructural de enzimas divergentes, pues son menos sensibles a la variabilidad que se pueda presentar entre las secuencias, como la ya mencionada presencia/ausencia de estructuras regulares e inclusive dominios estructurales completos (5, 6, 38, 50, 65, 66, 67, 73).

Cuando mapeamos los 49 Patrones Distintivos de Hidrofobicidad (PDH) en las ramas de nuestro dendograma estructural con las 172 secuencias positivas, pudimos apreciar cuáles resultaron ser las agrupaciones entre las distintas secuencias, a la vez que fuimos capaces de identificar la disposición peculiar de los 49 PDHs mapeados sobre él (Figura 8 y Tabla 3). Lo primero que podemos notar es que hay patrones distintivos de hidrofobicidad con una distribución filogenética más amplia. Esto es, el algoritmo de HCAmA encuentra que algunos de ellos son más recurrentes que otros [Figura 10]. Por ejemplo, el patrón *Tannerela* (recordemos que los patrones llevan el mismo nombre que las secuencias en las que fueron encontradas) que corresponde a la secuencia de *Tannerela sp.*, es el patrón más abundante, ya que se encuentra en 24 secuencias (o especies) distintas [Figura 9 y 10]. Esta secuencia/patrón está clasificada dentro del orden de las *Bacteroidetes*, que son el segundo grupo más representado en el dendograma.

Aunque analizar minuciosamente los diagramas de HCA para 172 secuencias (que ya incluyen a los 49 PDH) no figura dentro de nuestros objetivos, nos dimos a la tarea de hacer una inspección visual detallada de los mencionados diagramas en algunas de las secuencias (las de los patrones con mayor representación, así como algunos diagramas de secuencias no-repetidas que detallaremos más adelante).

Cuando observamos el patrón de *Tannerela sp.* en un diagrama de HCA [Figura 9], podemos observar que es una secuencia grande, con aprox. 460 aminoácidos (el tamaño promedio para otras Lisozimas en el análisis es de 350 aa´s). La presencia de una gran cantidad de agrupamientos de aminoácidos hidrofóbicos de gran tamaño, en la mayoría de los casos de más de 4 aminoácidos, que es el tamaño "estándar" para definir la longitud mínima de un agrupamiento o *cluster* que sea capaz de formar una estructura secundaria regular pequeña, sugieren un ordenamiento estructural "distintivo" para esta secuencia (6, 8, 65), en el que notamos una composición abundante y "regular" (que no se encuentran dispersos y/o a grandes distancias entre ellos) de agrupamientos de aminoácidos hidrofóbicos, al mismo tiempo que existen muy pocas regiones en las que los aminoácidos polares son predominantes, regiones normalmente reconocidas como "desordenadas" y con frecuencia asociadas a las interacciones entre proteínas (42, 65) [Figura 9].

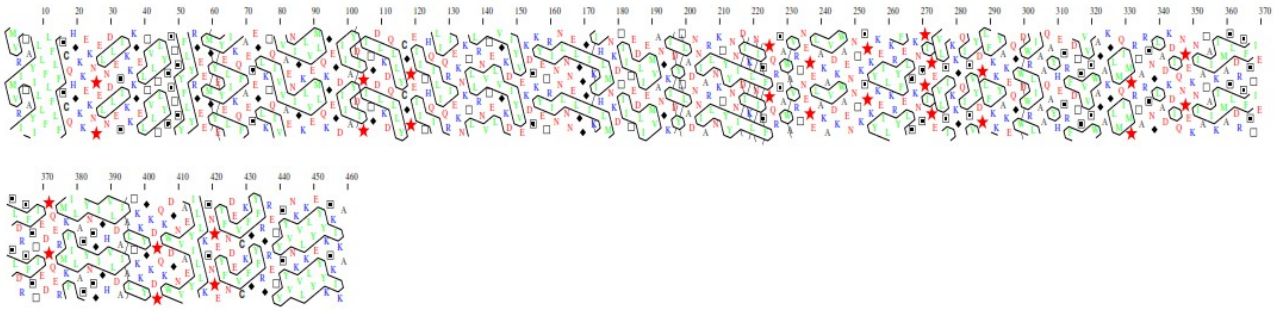

Figura 9. La secuencia de *Tannerella sp.*, que es la secuencia en la que se encontró el patrón más difundido, presente en otras 23 secuencias del total de las Lisozimas estudiadas.

Lamentablemente, el algoritmo HCAmA no indica gráficamente aquellos patrones distintivos que fueron identificados para determinada secuencia (esto es, no los resalta en la secuencia de la proteína ni en un diagrama de HCA), por lo que no podemos saber cuáles son los patrones que encuentra; sin embargo, nosotros consideramos que es posible que, dado el tamaño de la secuencia, la gran cantidad de agrupamientos hidrofóbicos que se encuentran y dad la evidente composición ordenada de la secuencia de *Tannerela sp.* (Figura 9), HCAmA esté identificando múltiples patrones PDH (dentro de la misma secuencia) que se encuentran en múltiples secuencias (del alineamiento completo); lo que de manera muy sencilla, explicaría por qué el patrón de *Tannerela sp.* se encuentra en 24 secuencias. Si estamos en lo correcto, esta secuencia es la más difundida porque en realidad posee múltiples patrones (¿mini-patrones?) que se encuentran en muchas otras secuencias. Hacemos notar que, contrario a lo que nuestra interpretación previa podría sugerirnos, pensamos que la frecuencia de un patrón no está en función de su tamaño, sino

de la composición única, en términos de composición/ubicación de agrupamientos hidrofóbicos, que las secuencias con patrones más prevalentes poseen, algo que nos parece evidente en el diagrama de HCA de la secuencia de *Tannerela sp.* [Figura 9].Estamos conscientes de que la única forma de corroborar la idea planteada en párrafos anteriores, es haciendo minuciosas comparaciones pareadas de los 34 PDH contra todas aquellas secuencias en las que se identificó dos o más patrones; sin embargo, por limitantes en cuanto al tiempo para llevar a cabo un análisis tan grande y minucioso, este acercamiento no fue llevado a cabo.

A pesar de lo anterior, para hacer un intento de verificar nuestras ideas en al menos otra secuencia/patrón más allá de *Tannerela sp.*; decidimos analizar someramente el diagrama de HCA con la secuencia del segundo patrón más frecuente, el de *Acidobacterium* (diagrama de HCA no mostrado) [Figura 10].

Esta secuencia/patrón, aunque es de una longitud menor a la media (sólo 250 aa´s), pensamos que debe su prevalencia a que (al igual que la secuencia/patrón de *Tannerela sp.*) posee una distribución peculiar de agrupamientos de aminoácidos hidrofóbicos ininterrumpidos; los cuales creemos que son los verdaderos culpables de la frecuencia de este patrón, más no el tamaño de la secuencia.

Dejando a un lado las limitantes técnicas, el dendograma estructural resume una gran cantidad de información de una forma muy amigable y conveniente [Figura 8], lo que nos permitió percatarnos de otro resultado interesante; y esto es que todas las secuencias presentan al menos un PDH, e inclusive, en algunos casos, presentan hasta tres PDH al mismo tiempo (Figura 8 y Tabla 3). Consideramos que esto no es un error ni un artilugio metodológico, pues, como ya mencionamos, los PDH corresponden a un acomodo espacial distintivo o “peculiar” de aminoácidos hidrofóbicos que posiblemente estén formando estructuras secundarias regulares o que sean parte de dominios estructurales/catalíticos completos (en ambos casos, dependiendo del tamaño podrían ser múltiples), como el dominio glucósido-hidrolasa (79) o el distintivo acomodo de aminoácidos hidrofóbicos que rodean un TIM-Barrel central que también se presenta en algunas otras Lisozimas (77, 66).

La segunda evidencia que apoya que esto no es un artilugio metodológico, es que en una muestra más diversa de proteínas (que incluía todas las secuencias con estructura reportada en el PDB hasta el 2004), en el trabajo de Pedro Silva se reporta, de igual manera, que una cantidad no especificada de secuencias analizadas son compatibles con múltiples PDH (5).

A pesar de que tenemos confianza en que todos los resultados obtenidos (descritos con las ideas que acabamos de plantear) poseen significado biológico (es decir, realmente reflejan cualidades intrínsecas a la estructura de las proteínas y la influencia que los aminoácidos hidrofóbicos poseen sobre ellas), no podemos omitir que, ciertamente, la notoria representatividad de sólo algunos de los patrones podría deberse únicamente por coincidencias azarosas. Si este es el caso, podríamos estar ante un escenario en el cual, el algoritmo HCAmA (y el fundamento teóricos en el que se basa) son insuficientes o carecen de la resolución/precisión necesaria para analizar la composición de las secuencias -que entonces no podrían considerarse más allá de falsos positivos- (6, 8); que es una de las advertencias que se hacen en el trabajo de Pedro Silva (5).
Hacemos énfasis en que nosotros analizamos secuencias de tamaño promedio, algunas de las cuales eran menores a 120 aa´s. Esto llama nuestra atención porque la advertencia que hace notar Pedro Silva va dirigida, principalmente, a la incertidumbre de analizar secuencias cortas (el propone, explícitamente, que analizar múltiples secuencias con menos de 80 aa´s podría ser problemático [5]).

Como lo mencionamos al principio de este trabajo, existen proteínas que no pueden ser analizadas de forma confiable mediante el algoritmo HCAmA, principalmente porque la composición de su secuencia posee muy pocos aminoácidos hidrofóbicos o por que los mismos no son fundamentales para la adquisición de una determinada estructura tridimensional, lo que podría indicar que, por ejemplo, en algunas proteínas, son otros los factores que ejercen una mayor influencia para la adopción de una estructura tridimensional particular; como serían las Proteínas asociadas a membrana y/o las que se encuentran en Matriz Extracelular, cuya estructura depende, principalmente, de enlaces disulfuro entre Cisteínas para mantener su integridad en ambientes altamente reductores (38).

Aun con todo lo anterior, seguimos pensando que la prevalencia y concordancia filogenética que los PDHs (en ocasiones múltiples y en la misma o diversas secuencias) nos muestran, resultan biológicamente significativa, debido a que:

1) Los posibles falsos positivos que pueden surgir durante los análisis comparativos manuales de los diagramas HCA se eliminaron desde los primeros pasos gracias a la automatización de las comparaciones de secuencia mediante el algoritmo HCAmA **2)** Este algoritmo cuenta con una función puntuadora que, precisamente, toma en cuenta la composición y el tamaño de las

secuencias comparadas para poder identificar los falsos positivos, minimizando así la identificación de resultados erróneos o poco significativos **3)** Durante la realización de este trabajo, se hicieron múltiples pruebas para verificar que el valor de penalización a la apertura de *gaps* favoreciera resultados estadísticamente significativos, al mismo tiempo que se mantenía baja la identificación e incorporación al análisis de posibles falsos positivos **4)** En la literatura, haciendo uso de otras aproximaciones estructurales, se ha demostrado que existen tendencias diferenciales a formar estructuras secundarias y/o dominios, conservados, los cuales, aquí analizamos en términos de Patrones Distintivos de Hidrofobicidad (PDH). Nuestro trabajo -en correspondencia con tales hallazgos-, demuestra que algunos (las estructuras/dominios y por ende, los patrones) son más frecuentes y prevalentes en una mayor cantidad de proteínas de diversos tipos, funciones, e incluso pertenecientes a organismos de distintos dominios; mientras que en otros casos, hay patrones poco difundidos que se ven sólo unas cuantas veces en proteínas con estructura muy particular o específica. Las secuencias de tales proteínas suelen ser referidas como "*Singlets*" [Figura 10] (77).

Siguiendo la idea anterior, para un total de 49 PDH, encontramos que 34 se encuentran en dos o más secuencias, mientras que 15 sólo se encuentran en la secuencia que les dio origen (Figura 10). Estas secuencias poseen una composición y agrupamiento distintivo de aminoácidos hidrofóbicos (a nivel secuencia) que no se encuentra en ninguna otra de las secuencias presentes en el alineamiento.

Ahora, si nos referimos a la Figura 10, podremos notar que, al menos 9 secuencias/patrón poseen una gran prevalencia (por ejemplo, el patrón "*Tannerela*" fue identificado en 24 secuencias portadoras), mientras que para 5 secuencias, la prevalencia es relativamente baja, en comparación. Nuevamente, resulta notorio que, como mencionamos, 15 PDH sean "*Singlets*".

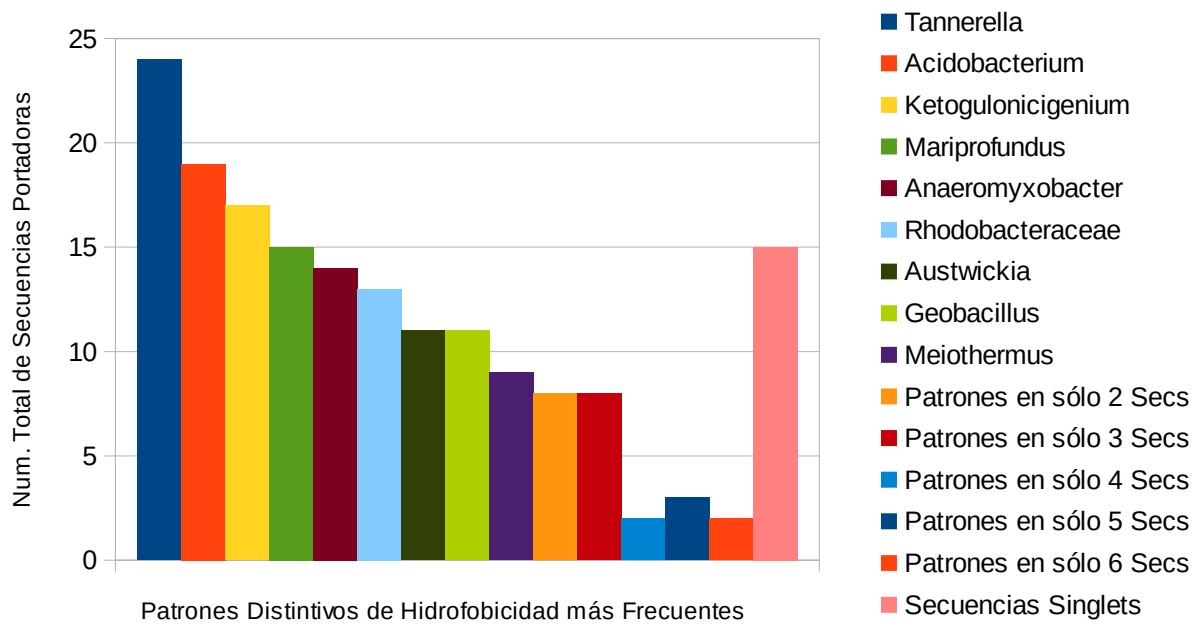

Figura 10. Tabla de frecuencias que muestra el número total de secuencias portadoras que poseen al menos uno de los 49 PDH.

Consideramos que es muy posible que los patrones que representan éstos *Singlets* no sean huérfanos (es decir, que no haya secuencias con características similares dentro de la familia de las Lisozimas). Resulta más plausible pensar que tal patrón no se encontró en nuestro alineamiento final (*dataset*), el cual fue sometido al análisis mediante el algoritmo HCAmA; por lo que, de incluir un número mayor de secuencias de Lisozima en el análisis, es posible que algunos de los patrones *Singlets* puedan ser identificados en alguna otra secuencia que ya se haya reportado.

Pensamos que, de nuevo, este resultado confirma la “tendencia” que se ha notado en una gran cantidad de análisis estructurales de diversas proteínas, y, en general este es que ciertos patrones/estructuras son más frecuentes y abundantes que otros (5, 6, 38, 65, 66, 67, 73, 77).

8. CONCLUSION

En este trabajo hemos abordado el estudio de las Lisozimas de Bacterias, que anteriormente, sólo se habían estudiado a una escala limitada. Resulta evidente que las bases de datos contienen muchas secuencias de tales Lisozimas; sin embargo, las mismas casi no se estudian, y pensamos que las principales razones por las que esto sucede es debido a que se trata de secuencias divergentes, variables y estadísticamente poco significativas. Además, con frecuencia no existe estructura tridimensional resuelta que permita hacer comparaciones a nivel estructural, lo que, en la mayoría de los casos, impide acercamientos a gran escala; algo que con la metodología que aquí presentamos, no implicaría una fuerte limitación para llevar a cabo un estudio.

La literatura tiene muchos ejemplos de metodologías (la mayoría de ellas complejas y demandantes) que tienen el propósito de identificar similitudes a nivel de secuencia o estructura tridimensional de diversas biomoléculas a partir de la comparación con otras ya conocidas. Esperamos que este texto refleje la noción de que, muy posiblemente, acercamientos como éste, en donde se hacen inferencias sobre la estructura de las proteínas partiendo sólo de las secuencias, sean el futuro de los análisis de estructura y/o plegamiento proteico.

Con la misma intención, abordando un estudio sencillo pero lo más completo posible de las Lisozimas de Bacterias, esperamos haber dejado implícita la practicidad y utilidad que la automatización mediante el algoritmo de HCA_Analyze_Multiple_Aligns (HCAmA) de los análisis conducidos bajo el marco teórico del método de HCA, resultan en una metodología complementaria que es eficiente y precisa para determinar similitudes estructurales, aún entre proteínas de secuencia divergente.

Pudimos corroborar (trabajando con Lisozimas de Bacterias) que la detección de similitud estructural puede ser especialmente complicada cuando se trabaja con secuencias divergentes y/o estructuras variables, ya que, en ocasiones, enzimas pertenecientes a la misma familia proteica no necesariamente poseen estructuras tridimensionales similares perfectamente súper-imponibles (5, 38, 73); así que la elección de cuál estructura o templado usaremos para inferir las propiedades tridimensionales de nuestra proteína/molécula problema no es un asunto trivial ni mucho menos sencillo. Complicando aún más las cosas, la mayoría de los análisis bioinformáticos que se hacen a nivel estructural, sólo toman en cuenta un dominio proteico a la vez, por lo que los estudios que

abordan la comparación de proteínas homo u heteromultiméricas (compuestas por repeticiones del mismo dominios o por múltiples dominios distintos, respectivamente) suelen ser muy limitados, pues por las restricciones técnicas que impone el modelado de un número muy grande de interacciones fisicoquímicas entre múltiples dominios, no es posible analizar biomoléculas multiméricas (38).

Enfatizamos que la metodología aquí descrita no posee las limitaciones arriba mencionadas, pues es factible trabajar con la totalidad de las secuencias, y es irrelevante sin con ellas se forman múltiples homo u heterodominos, de la misma forma en que el tamaño promedio de los mismos no es un factor limitante. Así, combinando un acercamiento que puede ser simultáneo para la comparación entre secuencias/estructuras completas, el método de HCA y su subsecuente automatización con el algoritmo HCAmA, favorecen el estudio de biomoléculas altamente divergentes que con frecuencia no se pueden trabajar (o el resultado es inconcluso) si el enfoque es unilateral; esto es, cuando los análisis son sólo de secuencia o estructura, y, esto sólo es posible si podemos asumir que se cuenta con una estructura resuelta para comparar.

Utilizar los Patrones Distintivos de Hidrofobicidad (PDH) nos ha permitido identificar aquellas secuencias de Lisozimas que serían buenas candidatas para estudios comparativos, ya sea empleando métodos bioinformáticos, normalmente enfocados a nivel de secuencia (más rápidos, sencillos y menos costosos) o experimentales (como la resolución de una estructura tridimensional mediante estudios cristalográficos). Aunque estamos conscientes de que el estudio se podría extender hacia el análisis pareado de los diagramas de HCA; la comparación mediante estos diagramas de HCA consume una cantidad de tiempo y esfuerzo significativo, y eso sin mencionar que las interpretaciones humanas suelen ser subjetivas (que es una de las críticas más arraigadas al método no-automatizado de HCA).
En este trabajo estamos exentos de la subjetividad de la interpretación de los diagramas de HCA gracias a la automatización de las comparaciones estructurales mediante el algoritmo HCAmA. Este acercamiento nos permitió identificar con confianza aquellas secuencias que muestran similitudes en cuanto a posición y composición de sus agrupamientos de aminoácidos hidrofóbicos, además de poder resumir toda esta información y plasmarla gráficamente (en un dendograma) para observar los agrupamientos estructurales que resultan de tales similitudes.

Además, esperamos que este trabajo aporte evidencia a favor para el uso generalizado del método de Análisis de los Agrupamientos (*Clusters*) Hidrofóbicos (HCA) por sus notorias ventajas sobre otras metodologías de análisis estructural; esto debido no sólo a su eficiencia, accesibilidad y valor, sino porque, en combinación con otras metodologías (Súper-Imposición de Estructura Tridimensional, Algoritmos de búsqueda de bases de datos, etc.) permite obtener resultados robustos y significativos que podrían facilitar la confirmación de hipótesis verificables y de gran valor.

Este último punto es de especial utilidad, pues no olvidemos que, actualmente, se necesitan probar hipótesis cada vez más complejas que a su vez dependen de un número mayor de evidencia (secuencias). Esto, claramente representa un problema, pero es uno que, según nuestra perspectiva, se puede resolver si se aplican metodologías comparativas que, aun haciendo análisis sobre sólo las secuencias, sean capaces de hacer inferencias respecto a las las similitudes a nivel estructural, justo como el método de HCA y su automatización con el algoritmo de HCAmA.

Finalmente, confirmamos que los Patrones de Hidrofobicidad son características de la secuencia/estructura de las proteínas que pueden resultar en información muy útil para efectuar diversos estudios, como las comparaciones entre secuencias/estructuras divergentes, identificar templados para estudios de predicción de estructura tridimensional y estudiar proteínas multidominio -sin la exclusión de información-, entre otros. Sin embargo, como en muchos otros aspectos del estudio de la estructura de las biomoléculas, entender el significado biológico de tales hallazgos aún es complicado.

Con un buen nivel de certeza, sabemos que, individualmente, éstos patrones representan “tendencias” a formar determinada estructura secundaria, lo que a su vez se refleja en la adopción de un plegamiento tridimensional característico; pero, ¿qué influencia poseen estas características intrínsecas a las proteínas en fenómenos que poseen genuina importancia biológica?, como por ejemplo: ¿los patrones distintivos de hidrofobicidad obedecen a presiones de selección?, ¿nos sirven para clasificar proteínas de acuerdo al género del organismo portador?, ¿los patrones son reflejo de adaptaciones funcionales distintivas?., etc.

Por el momento no es posible contestar tales preguntas, sin embargo, las comparaciones entre muestras significativamente más grandes de secuencias para trabajar y algoritmos capaces de lidiar con un número mayor de variables dependientes (variables cuya probabilidad depende de la

ocurrencia de fenómenos controlados por alguna otra variable), ciertamente podrían acercarnos hacia la resolución de tales cuestiones. De lograr tales avances, muy posiblemente tendríamos acceso al análisis de las interacciones “globales” a distintos niveles de organización; análisis que no sólo resultarían muy codiciados en el estudio de la estructura de proteínas (pensemos en el impacto que esto tendría en otras ramas de la Biología, como la Ecología), lo que inmediatamente se vería reflejado en un aumento en la capacidad de predicción, identificación y comparación de la estructura de diversas proteínas, que en la muy personal opinión de este autor, engloba el objetivo “final” que todos los que llevamos a cabo estos estudios eventualmente tenemos como meta.

9. PERSPECTIVAS

1) En párrafos anteriores discutimos la presencia de secuencias que llamamos "*Singlets*", que son secuencias cuyos patrones de hidrofobicidad no pudieron ser identificados en ninguna otra secuencia de nuestro alineamiento final (*dataset*). Muy posiblemente, estos resultados reflejan las limitaciones de nuestra muestra, por lo que no los consideramos ni errores ni artilugios metodológicos. Nosotros hipotetizamos que tales patrones podrían encontrarse en otras secuencias de Lisozimas. Para confirmarlo, se podría reunir una muestra mayor de Lisozimas en la que se incluya la secuencia de, al menos, una representante de cada una de las 100 familias de glucósido-hidrolasas. Si los análisis se repiten con una muestra mayor, creemos que la proporción de *Singlets* podría disminuir significativamente.

2) Un acercamiento que nos gustaría efectuar en un futuro, es el de llevar a cabo la comparación minuciosa, mediante los diagramas de HCA, de los 34 patrones que están presentes en al menos dos o más secuencias. Anticipamos que las comparaciones pareadas a nivel estructural podrían ayudar a comprender mejor la relación que existe entre secuencia y estructura de éstas enzimas.

3) Para extender los alcances predictivos de este trabajo, proponemos reunir una muestra mayor de Lisozimas presentes en diversos organismos de diversos grupos taxonómicos (Virus, Bacterias, Animales, Plantas, Insectos, etc.) y determinar si existen patrones comunes que sean correspondientes con la diversificación de éstos organismos; esto es, que pudiésemos encontrar (si es que existe) un patrón que sea característico de los Virus, uno de Bacterias, otro de Eucariontes, etc. Efectuando tales análisis, creemos que sería factible comprobar si existe o no un "foldoma" característico de cada grupo taxonómico; al menos para el caso de las Lisozimas.

4) Otra alternativa para extender el trabajo, sería aquella que se enfocara en afrontar el problema del origen de éstas enzimas. Si bien tal cuestión es difícil de confirmar -y por la enorme cantidad de trabajo, el problema sólo se podría abordar a una escala menor-, proponemos hacer un análisis pareado de secuencias con el algoritmo HCAmA de secuencias de Lisozimas de fagos (las que se encuentren en las bases de datos) contra representantes clave de Eucariontes, Bacterias e Insectos. Las comparaciones pareadas nos permitirían identificar, uno a uno, aquellos casos en los que exista similitud estructural significativa entre las secuencias de Lisozimas que pertenecen a organismos de distintos reinos. Los resultados podrían contribuir con más evidencia para,

eventualmente, tratar de abordar de una forma más concreta, la resolución de la gran pregunta que hasta hoy, a pesar de numerosos intentos, sigue sin resolverse: ¿son todas las Lisozimas homólogas?, de ser así, ¿esto es válido aun para las que pertenecen a organismos de distintos Reinos?, y, si es así, ¿cuáles son las más conservadas, las de Procariontes, Eucariontes o Fagos?.

5) En aplicaciones que van más allá de las metas establecidas en este trabajo, proponemos que una metodología similar a la abordada por Pedro Silva (5), que consiste en analizar bases de datos completas en busca de Patrones Distintivos de Hidrofobicidad de proteínas de diversos tipos, y con ello, hacer estudios ya no sólo de tipo estructural, sino también funcionales, que nos permitan asociar la distribución, composición y frecuencia de los PDH con diversos aspectos biológicos, como por ejemplo, ¿las isomerasas poseen PDH característicos que son compatibles con otro tipo de enzimas, como, Liasas o Hidrolasas? O la distribución de aminoácidos hidrofóbicos es única para las enzimas representadas en esta categoría?, ¿Qué importancia tendrían los PDH para la función de éstas enzimas?. Acercamientos como el que acabamos de describir, creemos que tendrían una importancia fundamental para continuar con los estudios a escalas mayores en los trabajos de estructura de proteínas.

Finalmente, nos gustaría aplicar el marco teórico del método de HCA en conjunción con algún algoritmo que nos permita hacer comparaciones a gran escala (recordemos que en este trabajo utilizamos HCAmA, pero en la literatura se ha publicado, reportado y puesto a disposición de los usuarios uno nuevo, que toma en cuenta las regiones desordenadas de las proteínas) al estudio de los DUFs. Los *DUFs*, de las siglas en inglés *Domains of Unknown Function*, son precisamente esto, Dominios cuya función es desconocida, aunque poseen una gran representatividad y frecuencia en las proteínas de diversos grupos taxonómicos, principalmente Bacterias. Nuestra intención sería abordar el estudio de los DUFs para identificar los Patrones Distintivos de Hidrofobicidad que poseen, y con ellos hacer búsquedas en bases de datos (mediante procedimientos como HMMER o BLAST) para tratar de relacionarlos con los Dominios Estructurales o Funcionales ya anotados que se encuentren en las mismas.

10. GLOSARIO

Agrupamiento Hidrofóbico: Véase *Cluster* Hidrofóbico.

Algoritmo Puntuador: Es parte del algoritmo HCAmA y es el que se encarga de la delimitación de secuencias aceptadas positivas. Este algoritmo evalúa el cociente que existe entre la longitud de las secuencias, la composición de las mismas (en términos de aminoácidos hidrofóbicos) y la expectativa de obtener coincidencias azarosas. Si dicho cociente no es positivo, el algoritmo las considera como Falsos Positivos.

Aminoácido No Polar: Denota a los Aminoácido Hidrofóbicos en términos de la ausencia de carga neta. El método HCA sólo considera a V,I,L,F,M,Y,W para efectuar los análisis.

Aminoácido Polar: Aminoácidos Cargados o no Cargados, Positivos, Neutros o Negativos que son Hidrofílicos.

***Cluster* Hidrofóbico:** Agrupamiento de aminoácidos hidrofóbicos (bajo consideraciones especiales, un sólo aminoácido puede ser considerado como parte de un mismo *cluster* o agrupamiento) que sirven de unidad de estudio/análisis para el método HCA y el algoritmo HCAmA.

***Dataset*:** También llamado "alineamiento final"; corresponde al conjunto de secuencias alineadas (con la Matriz HCA) en formato *.pir* que sirven de archivo de entrada para el algoritmo HCAmA.

Distancia de Conectividad: Es la distancia (aproximadamente de 4 aminoácidos) que nos ayuda a identificar cuando, un aminoácido hidrofóbico "*i*" (localizado en algún punto espacial de un diagrama de HCA) pertenece al mismo o a distintos agrupamientos de aminoácidos hidrofóbicos.

Dominio: Es una unidad funcional o estructural distintiva componente de la estructura tridimensional de una proteína. Los dominios tienen la propiedad de plegarse de forma independientemente al resto de la estructura. Una proteína puede estar constituída por uno o múltiples dominios (multidominio); de los que puede existir repeticiones del mismo (homodominios) o todos y cada uno de los dominios presentes, ser distinto a los demás (heterodominio)

E-value: Abreviatura en inglés a “valor de expectativa” que indica el número de coincidencias que se esperan debido simplemente al azar cuando se utiliza una secuencia templado (*Query*) contra una base de datos. En este trabajo, el algoritmo HCAmA ajusta el valor de “E” de acuerdo a la cantidad de aminoácidos hidrofóbicos promedio que poseen las secuencias que conforman el alineamiento final o *dataset*.

Estructura Secundaria Regular: Es la denominación que se emplea para referirse a Hojas-Beta (paralelas o antiparalelas), Alfa-Hélices y estructuras de inteconexión, como “asas” o *loops;* aunque también hay estructuras secundarias regulares más pequeñas, como las Vueltas-Beta, etc.

Falso Positivo: Es el escenario en el cual el algoritmo HCAmA consideraría como significativa la comparación de la secuencia “X” contra la “Y”, cuando en realidad, no existe tal significancia, o la misma es producto de artilugios metodológicos.

Falso Negativo: Es el escenario en el cual el algoritmo HCAmA omitiría la significancia al comparar la secuencia “X” contra la “Y”, a pesar de que, en realidad, la comparación minuciosa de ambas (utilizando otro método) podría arrojar coincidencias previamente omitidas.

Familia Proteica: Grupo de proteínas que comparten un origen común.

Foldoma: Se define como la totalidad de los plegamientos que presentan las diversas estructuras de las proteínas; sin embargo, en este trabajo reducimos el alcance del término, y lo consideramos como el conjunto de plegamientos que definen (o que podrían identificar) sólo a las enzimas Lisozimas y a los plegamientos que les definen dentro de una misma categoría.

Gap: Es un “espacio” o “hueco” introducido en una secuencia para maximizar su parecido con otra. Los *Gaps* simulan eventos de Inserción-Deleción de nucleótidos/aminoácidos que pueden dificultar el alineamiento de regiones de similitud entre secuencias.

HCAmA: Abreviación de *Hydrophobic Cluster Analysis Multiple Aligns*, Es el algoritmo con el que se evalúan las secuencias en términos numéricos/estadísticos de acuerdo a la longitud y composición de aminoácidos hidrofóbicos. En otras palabras, es el algoritmo que analiza y evalúa los agrupamientos de aminoácidos hidrofóbicos.

Indel: Es la abreviación para Inserciones-Deleciones. Los Indels son producto de la divergencia natural que ocurre entre secuencias homólogas (mutaciones) durante eventos de especiación. En los alineamientos de secuencias, los Indels son "compensados" haciendo uso de "espacios" o "gaps"; lo que representa la ausencia (e incluso ganancia) de segmentos que no están presentes en todas las secuencias comparadas.

Loop: Los *loops* o "asas" son estructuras secundarias comunes que normalmente sirven como estructuras de interconexión entre distintas estructuras secundarias regulares (alfa-hélices y hojas-beta) e incluso, de dominios estructurales/catalíticos completos. Son regiones compuestas de abundantes aminoácidos polares (hidrofílicos), normalmente desordenadas y tolerantes a los *Indels* que ocurren de forma natural en las secuencias.

Ortólogo: Genes o secuencias homólogas en distintas especies que son producto de eventos de especiación.

Parálogo: Genes o secuencias producto de la duplicación de los Genes dentro del genoma de la misma especie.

Patrón Distintivo de Hidrofobicidad: Corresponde a secuencias que poseen un porcentaje de similitud de HCA menor al 60% (el *HCA_Score*) en comparación con todas las demás secuencias que se encuentran presentes en el alineamiento analizado mediante HCAmA.

PDH: Abreviatura de Patrón Distintivo de Hidrofobicidad.

Penalización: Es la puntuación negativa que recibe la inserción de demasiados *Gaps* en un alineamiento cualquiera. Dadas presiones de selección distintivas, usualmente se da un tratamiento diferencial a los *Gaps* internos que a los externos (regiones N y C de las secuencias).

Plegamiento: Es la descripción de la conformación estructural que adoptan los dominios en la estructura tridimensional de alguna proteína. El plegamiento permite delimitar la disposición espacial de las estructuras secundarias regulares que conforman los dominios estructurales/funcionales.

Secuencias Aceptadas o Positivas: Secuencias que son consideradas como significativas para llevar a cabo comparaciones múltiples. En el archivo de salida que arroja el algoritmo HCAmA, se define explícitamente (con la notificación "*above threshold*") aquellas secuencias que cumplen con tales características.

***Singlets*:** Secuencia cuyo Patrón Distintivo de Hidrofobicidad no se identificó en ninguna otra secuencia del alineamiento final. Son secuencias con un Patrón único no compartido.

***Super-Folds o* Súper-Plegamientos.** Definidos en el trabajo de Orengo *et al*, en 1994 (77), se consideran como plegamientos tan comunes que son capaces de describir el 30% de todas las proteínas globulares. El TIM-Barrel y el Rosmann Fold son ejemplos de súper-plegamientos.

Tendencia: Predisposición a formar estructuras regulares (ya sea secundarias e incluso terciarias).

TIM-Barrel: Definido como un Súper-Plegamiento en el trabajo de Orengo *et al*, en 1994, por ser una estructura tridimensional terciaria sumamente común en las proteínas globulares. Es una estructura con forma de barril compuesto de Triosa Fosfato Isomerasa, que consiste de ocho hojas-beta paralelas, en las que cada par de hojas adjacentes están conectadas mediante un *loop* que a su vez contiene un alfa-hélice.

Xenólogo: Genes o secuencias usualmente homólogas adquiridos mediante Transferencia Horizontal de Genes. Puede ser un proceso que se de a nivel intra e interespecies.

11. REFERENCIAS

1.- Becerra A, Islas S, Leguina JI, Silva E, Lazcano A. 1997. Polyphyletic Gene Losses Can Bias Backtrack Characterizations of the Cenancestor. J Mol Evol. 45(2):115-8.

2.- Bausch-Jurken MT, Mahnke-Zizelman DK, Morisaki T, Sabina RL. 1992. Molecular Cloning of AMP Deaminase Isoform LJ Biol Chem. 267(31):22407-13.

3.- Marotta R, Parry BR, Shain DH. 2009 Divergence of AMP Deaminase in the Ice Worm Mesenchytraeus solifugus (Annelida, Clitellata, Enchytraeidae) Int J Evol Biol. 2010:715086.

4.- Gaboriaud C, Bissery V, Benchetrit T, Mornon JP. 1987. Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences. FEBS Lett. 1224(1):149-55.

5.- Silva PJ. 2008. Assessing the reliability of sequence similarities detected through hydrophobic cluster analysis. Proteins. 70(4):1588-94.

6.- Lemesle-Varloot L, Henrissat B, Gaboriaud C, Bissery V, Morgat A, Mornon JP. 1990. Hydrophobic cluster analysis: procedures to derive structural and functional information from 2-D-representation of protein sequences. Biochimie. 72(8):555-74.

7.- Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. Nat Rev Genet. 31;12(10):692-702.

8.- Callebaut I, Labesse G, Durand P, Poupon A, Canard L, Chomilier J, Henrissat B, Mornon JP. 1997. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. Cell Mol Life Sci.; 53(8):621-45.

9.- Ogasawara, N., Goto, H., Yamada, Y., Watanabe, T., and Asano, T. 1982. Biochim. Biophys. Acta 714,298-306

10.- Yun, S.L., & Suelter, C. H. 1978. J. Biol. Chem. 253,404-408

11.- Ogasawara, N., Goto, H., Yamada, Y., and Watanabe, T. 1984. Int. J. Biochem. 16,269-273

12.- Morisaki, T., Gross, M., Morisaki, H., Pongratz, D., Zollner, N., and Holmes, E. W. 1992. Proc. Natl. Acad. Sci. U. S. A. 89, 6457-6461

13.- Altschul, S. F., et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acid Res.

14.- Thompson JD, et al. 1997. "The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools". Nucleic Acids Research 25 (24): 4876–4882.

15.- Gutiérrez-Preciado, A., Romero, H. & Peimbert, M. (2010). An Evolutionary Perspective on Amino Acids. Nature Education3(9):29

16.- Irwin, D. et al. 2011. Evolution of the mammalian lysozyme gene family. BMC Evolutionary Biology, 11:166

17.- Callewaert, L y Michiels CW. 2010. Lysozimes in the animal kingdom. J. Biosci. 35(1), 127-160

18.- Eugene V. Koonin and Michael Y. Galperin. 2003. Sequence Evolution – Function. Computational Approaches in Comparative Genomics. National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health. Boston Kluwer Academic; ISBN-10: 1-40207-274-0 http://www.ncbi.nlm.nih.gov/books/NBK20260/

19.- Henrissat B, Callebaut I, Mornon JP, Fabrega S, Lehn P, Davies G. 1995. "Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases". Proc. Natl. Acad. Sci. U.S.A. 92 (15): 7090–7094.

20.- Henrissat B, Davies G. 1995. "Structures and mechanisms of glycosyl hydrolases". Structure 3 (9): 853–859.

21.-Magrane M. and the UniProt consortium. 2011. UniProt Knowledgebase: a hub of integrated protein data. Database: bar009. Swiss-Prot http://www.ebi.ac.uk/uniprot/

22.- Finn, RD. Clements, J. y Eddy, SR. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Research. Web Server Issue 39:W29-W37.

23.- Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, and Henrissat B. 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res. 37:D233-8. CAZypedia. http://www.cazypedia.org/index.php/Main_Page

24.- Shewale JG, Sinha SK, Brew K. 1984. "Evolution of alpha-lactalbumins. The complete amino acid sequence of the alpha-lactalbumin from a marsupial (Macropus rufogriseus) and corrections to regions of sequence in bovine and goat alpha-lactalbumins". J. Biol. Chem. 259 (8): 4947–56.

25.- Hall L, Campbell PN. 1986. "Alpha-lactalbumin and related proteins: a versatile gene family with an interesting parentage". Essays Biochem. 22: 1–26.

26.- Irwin DM, Wilson AC. 1989. "Multiple cDNA sequences and the evolution of bovine stomach lysozyme". J. Biol. Chem. 264 (19): 11387–93.

27.- Kamei K, Hara S, Ikenaka T, Murao S. 1988. "Amino acid sequence of a lysozyme (B-enzyme) from Bacillus subtilis YT-25". J. Biochem. 104 (5): 832–6.

28.- Nitta K, Sugai S. 1989. "The evolution of lysozyme and alpha-lactalbumin". Eur. J. Biochem. 182 (1): 111–8.

29.- Stuart DI, Acharya KR, Walker NP, Smith SG, Lewis M, Phillips DC. 1986. "Alpha-lactalbumin possesses a novel calcium binding loop". Nature 324 (6092): 84–7. doi:10.1038/324084a0.

30.- Nitta K, Tsuge H, Sugai S, Shimazaki K. 1987. "The calcium-binding property of equine lysozyme". FEBS Lett. 223 (2): 405–8. doi:10.1016/0014-5793(87)80328-9.

31- Vollmer, W. *et al.* 1997. Pesticin displays muramidase activity. J Bacteriol. 179(5): 1580–1583.

32.- Protein Data Bank. http://www.rcsb.org/pdb/home/home.do

33.- Friend, an integrated analytical front-end application for bioinformatics. Bioinformatics. 2005 Sep 15;21(18):3677-8. Epub 2005 Aug 2. PubMed

34.- Gedit. http://projects.gnome.org/gedit/

35.- DALI. http://ekhidna.biocenter.helsinki.fi/dali_server/

36.- MAMOOTH. https://ub.cbm.uam.es/software/online/mammoth.php

37.- Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. Science. 278(5338):631-7.

38.- Petsko, Gregory A. & Dagmar Ringe. Protein Structure and Function. London: New Science, 2004. Print.

39.- Krebs, J.E., Goldstein, E.S. y Kilpatrick, S.T. 2011. Lewin's Genes X. Sudbury, MA, USA: Jones y Bartlett Learning

40.- Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., y Walter, P. 2008. Molecular Biology of the Cell. 5'th ed. Garland Science Publishing, London.

41.- Hasegawa H, Holm L. 2009. Advances and pitfalls of protein structural alignment. Curr Opin Struct Biol;19(3):341-8.

42.- Eudes R. *et al.* 2007. A generalized analysis of hydrophobic and loop clusters within globular protein sequences. BMC Structural Biology.

43.- Gibas, C. & Jambeck, P. 2001. Developing Bioinformatics computer skills. O`Reilly.

44.- Chanma, C. *et al.* 20010. An Algorithm to Compute protein homology Based on Hydrophobic cluster analisis. (sin revista).

45.- Cuff A.L. *et al.* 2013. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. Nucleic Acids Res.

46.- Kufareva I, Abagyan R. 2012. Methods of protein structure comparison. Methods Mol Biol. 2012;857:231-57.

47.- Taylor W.R., Orengo C.A. Protein structure alignment. J. Mol. Biol. 1989;208:1-22.

48.- Taylor W.R., *et al.* Protein structure: geometry, topology and classification. Rep. Prog. Phys. 2001;64:517-590.

49.-. Nitta K1, Sugai S. 1989. The evolution of lysozyme and alpha-lactalbumin. Eur J Biochem. 1;182(1):111-8.

50.- Baussand J, Deremble C, Carbone A. 2007. Periodic distributions of hydrophobic amino acids allows the definition of fundamental building blocks to align distantly related proteins. Proteins. 67(3):695-708.

51.- Murat D, Byrne M & Komeili A. 2010. Cell Biology of Prokaryotic Organelles. Cold Spring Harb. Perspect Biol. 10:a000422.

52.- Baker, D & Sali, A. 2001. Protein Structure Prediction and Structural Genomics. Science. 294 (5540), 93-96.

53.- Jespers L, Sonveaux E, Fastrez J. 1992. Is the bacteriophage lambda lysozyme an evolutionary link or a hybrid between the C and V-type lysozymes? Homology analysis and detection of the catalytic amino acid residues. J Mol Biol. 228(2):529-38.

54.- Patzer SI, Albrecht R, Braun V, Zeth K. 2012. Structural and mechanistic studies of pesticin, a bacterial homolog of phage lysozymes. J Biol Chem. 287(28):23381-96.

55.- Mizuguchi K, Deane CM, Blundell TL, Overington JP. 1998. HOMSTRAD: a database of protein structure alignments for homologous families. Protein Science 7:2469-2471.
56.- Kanehisa M., Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28(1):27-30.
57.- Grant A, Lee D, & Orengo C. 2004. Progress towards mapping the universe of protein folds. Genome Biology, 5:107
58.- Kyte, J. Doolittle, R.F. 1982. "A simple method for displaying the hydropathic character of a protein". J. Mol. Biol. 157 (1): 105–32.
59.- Rose, G. D. & Roy, S. 1980. Hydrophobic basis of packing in globular proteins. Proc Natl Acad Sci USA. 77(8): 4643–4647. PMCID: PMC349901
60.- Dworkin, J. E., & Rose, G. D. 1987. Hydrophobicity profiles revisited, in: Methods in Protein Sequence Analysis (K. A. Walsh, ed.), Humana Press, Clinton, NJ.
61.- The UniProt Consortium. 2013. Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Res. 41: D43-D47.
62.- Murzin A. G., Brenner S. E., Hubbard T., Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences andstructures. J. Mol. Biol. 247, 536-540.
63.- Flicek, P. et al. 2013. Ensembl. 2013. Nucleic Acids Research. 2013 41 Database issue: D48-D55.
64.- Dayhoff MO, Barker WC, Hunt LT. 1983. Establishing homologies in protein sequences. Methods Enzymol. 91:524-45.
65.-Faure G, Callebaut I. 2013. Comprehensive Repertoire of Foldable Regions within Whole Genomes. PLoS Comput Biol 9(10): e1003280.
66.- Henrissat B, Callebaut I, Fabrega S, Lehn P, Mornon JP, Davies G. 1996. Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases. Proc Natl Acad Sci U S A. 28;93(11):5674.
67.- Henrissat B, Davies G. 1997. Structures and mechanisms of glycosyl hydrolases. Curr. Opin. Struct. Biol. 1997 Oct;7(5):637-44.
68.- Thompson JD. *et al* 1997. "The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools". Nucleic Acids Research 25 (24): 4876–4882.
69.- Tamura K., Stecher G., Peterson D., Filipski A., Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Molecular Biology and Evolution30: 2725-2729.
70.- Saitou N. & Nei M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4:406-425.
71.- Baldauf SL. Trends Genet. 2003. Phylogeny for the faint of heart: a tutorial. (6):345-51.
72.- Balaji SI, Srinivasan N. 2001. Use of a database of structural alignments and phylogenetic trees in investigating the relationship between sequence and structural variability among homologous proteins. Protein Eng. 14(4):219-26.

73.- Patthy, László. Protein Evolution. Oxford: Blackwell Science, 1999.

75.- Weaver LH, Grütter MG, Remington SJ, Gray TM, Isaacs NW, Matthews BW. 1985. Comparison of goose-type, chicken-type, and phage-type lysozymes illustrates the changes that occur in both amino acid sequence and three-dimensional structure during evolution. J Mol Evol. (2):97-111.

76.- Balaji S, Srinivasan N.. 2001. Use of a database of structural alignments and phylogenetic trees in investigating the relationship between sequence and structural variability among homologous proteins. Protein Eng. (4):219-26.

77.- Orengo C.A. Jones D.T, Thornton J.M. 1994. Protein superfamilies and domain superfolds. Nature. 372(6507):631-4.

78.- Krystek SR Jr, Metzler WJ, Novotny J. 2001. Hydrophobicity profiles for protein sequence analysis. Curr Protoc Protein Sci.

79.- Koonin E. & Rudd K.E. A conserved domain in putative bacterial and bacteriophage transglycosylases. Trends Biochem Sci. 1994 Mar;19(3):106-7.

80.- (a) Alland C, *et al*. 2005. RPBS: a web resource for structural bioinformatics. Nucleic Acids Res. 33:W44-9.

80.- (b) Néron B. *et al*. 2009. Mobyle: a new full web bioinformatics framework. Bioinformatics. (22):3005-11. Epub. Portal en Línea: HCA-Draw [@rpbs 1.0.2]. [http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::rpbs.HCA]

81.- Brocchieri L. 2001. Phylogenetic inferences from molecular sequences: review and critique. Theor Popul Biol. (1):27-40.

82.- Michael S.Y. Lee. 2001. Unalignable sequences and molecular evolution. Trends in Ecology & Evolution, Volume 17, Issue 2.

83.- Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nature Methods 9(8), 772.

84.- Guindon, S. & Gascuel, O. (2003). A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood". Systematic Biology 52: 696-704.

85.- Goodacrea, N.F., Gerloffb, D.L. Uetzc, P. 2013. Protein Domains of Unknown Function Are Essential in Bacteria. mBio vol. 5 no.1 e00744-13.

****

12. ANEXOS

El dendograma estructural fragmentado para apreciarlo a detalle.

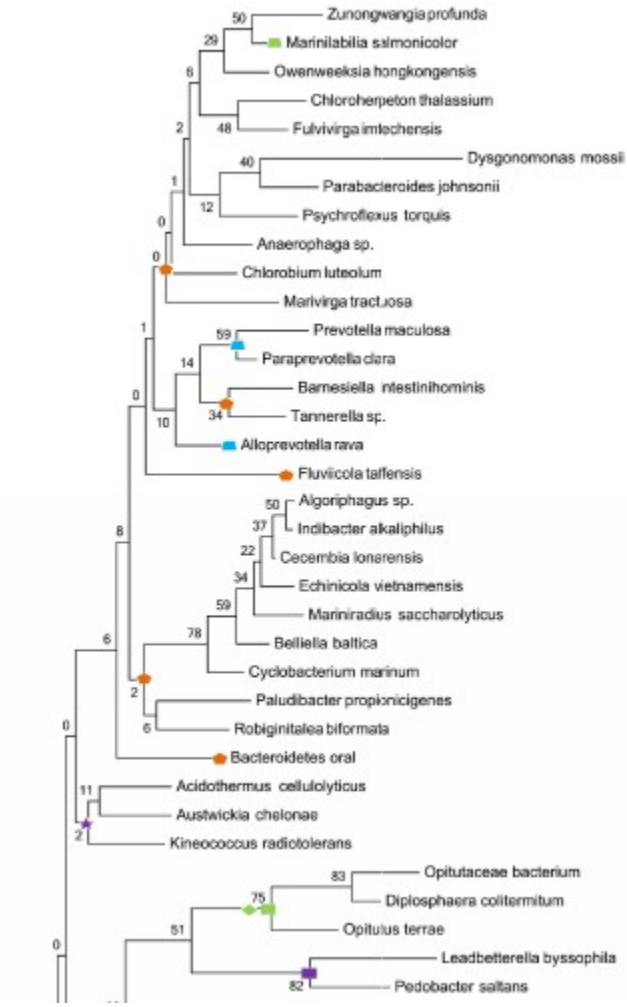

Figura 8. Dendograma estructural. Podemos notar que el dendograma consta de: 90 secuencias de *Proteobacterias*, 29 secuencias de *Bacteroidetes*, 16 secuencias de *Actinobacteria*, 19 secuencias de *Firmicutes* y 18 secuencias de algunos otros órdenes representados en menor proporción, como, *Verrucomicrobia* (5), *Synergistetes* (5) y *Deinococcus* (4), otros órdenes, como *Fusobacteria* y *Planctomycetes* sólo tienen una secuencia representante. El árbol también posee una única secuencia del Fago de *Escherichia coli*.

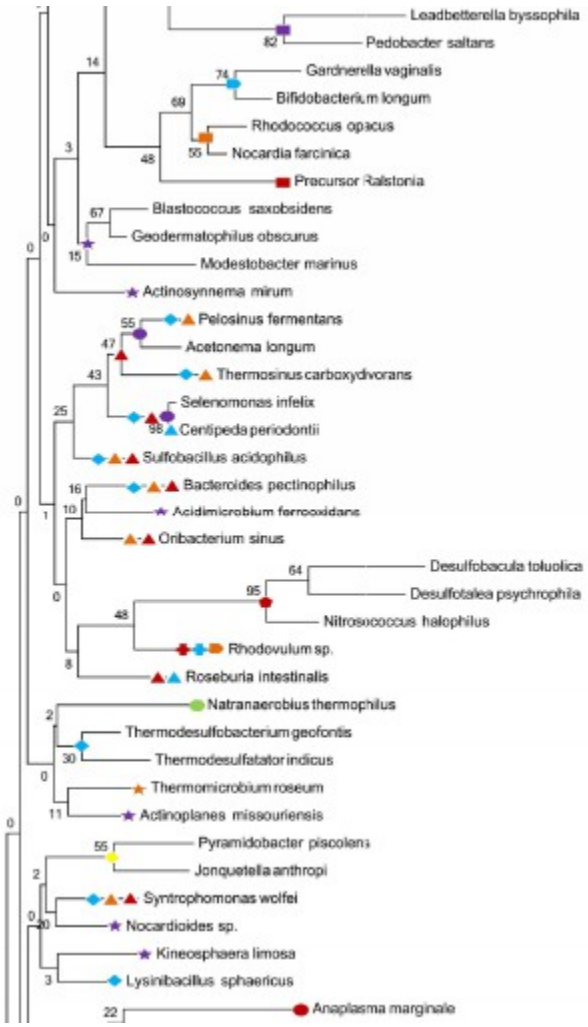

Figura 8. Dendograma estructural. Podemos notar que el dendograma consta de: 90 secuencias de *Proteobacterias*, 29 secuencias de *Bacteroidetes*, 16 secuencias de *Actinobacteria*, 19 secuencias de *Firmicutes* y 18 secuencias de algunos otros órdenes representados en menor proporción, como, *Verrucomicrobia* (5), *Synergistetes* (5) y *Deinococcus* (4), otros órdenes, como *Fusobacteria* y *Planctomycetes* sólo tienen una secuencia representante. El árbol también posee una única secuencia del Fago de *Escherichia coli*.

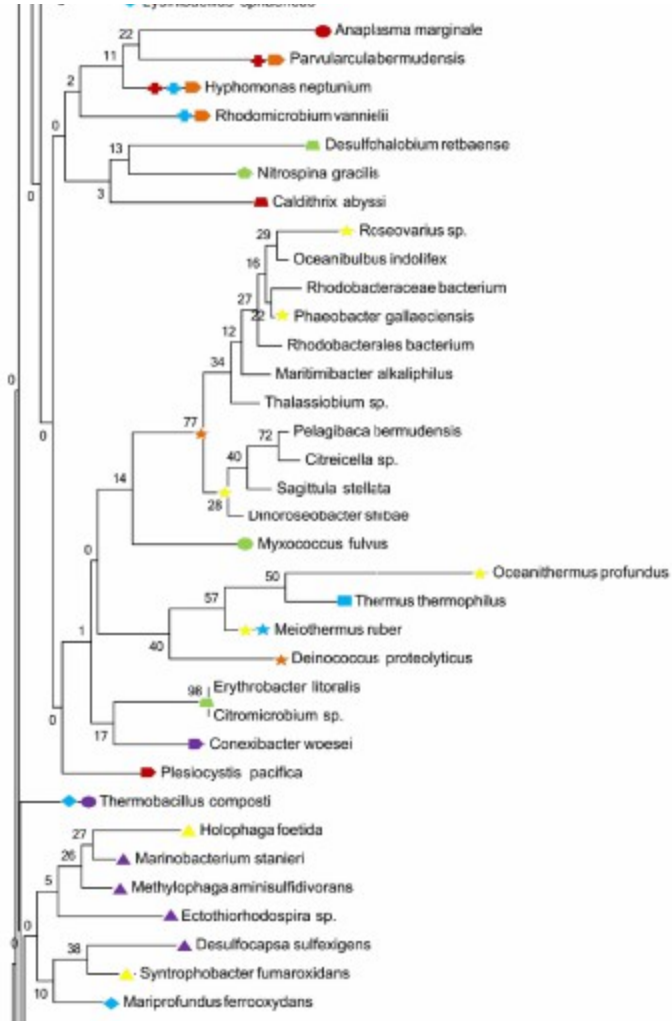

Figura 8. Dendograma estructural. Podemos notar que el dendograma consta de: 90 secuencias de *Proteobacterias*, 29 secuencias de *Bacteroidetes*, 16 secuencias de *Actinobacteria*, 19 secuencias de *Firmicutes* y 18 secuencias de algunos otros órdenes representados en menor proporción, como, *Verrucomicrobia* (5), *Synergistetes* (5) y *Deinococcus* (4), otros órdenes, como *Fusobacteria* y *Planctomycetes* sólo tienen una secuencia representante. El árbol también posee una única secuencia del Fago de *Escherichia coli*.

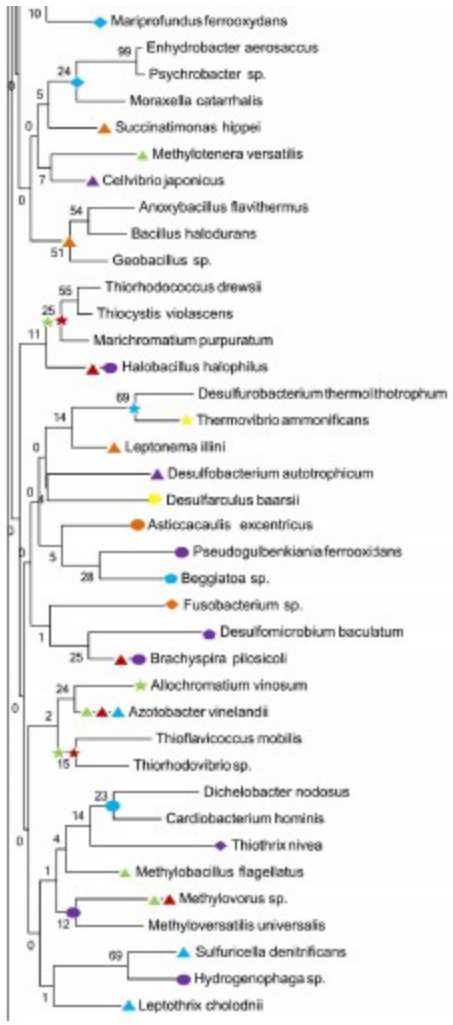

Figura 8. Dendograma estructural. Podemos notar que el dendograma consta de: 90 secuencias de *Proteobacterias*, 29 secuencias de *Bacteroidetes*, 16 secuencias de *Actinobacteria*, 19 secuencias de *Firmicutes* y 18 secuencias de algunos otros órdenes representados en menor proporción, como, *Verrucomicrobia* (5), *Synergistetes* (5) y *Deinococcus* (4), otros órdenes, como *Fusobacteria* y *Planctomycetes* sólo tienen una secuencia representante. El árbol también posee una única secuencia del Fago de *Escherichia coli*.

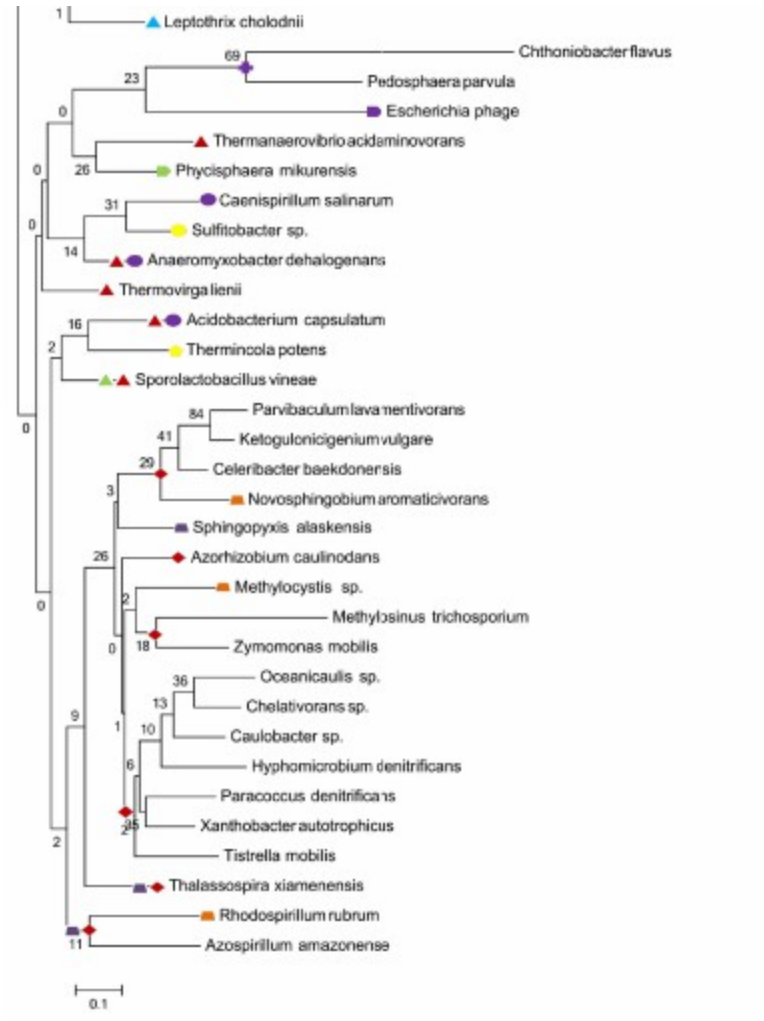

Figura 8. Dendograma estructural. Podemos notar que el dendograma consta de: 90 secuencias de *Proteobacterias*, 29 secuencias de *Bacteroidetes*, 16 secuencias de *Actinobacteria*, 19 secuencias de *Firmicutes* y 18 secuencias de algunos otros órdenes representados en menor proporción, como, *Verrucomicrobia* (5), *Synergistetes* (5) y *Deinococcus* (4), otros órdenes, como *Fusobacteria* y *Planctomycetes* sólo tienen una secuencia representante. El árbol también posee una única secuencia del Fago de *Escherichia coli*.

Tabla 3. Listado de los 49 Patrones Distintivos de Hidrofobicidad (PDH). En estas columnas, se muestran el símbolo y color de cada uno de los 49 PDH, así como su nombre y representatividad. La representatividad de los PDH (tercera columna) hace referencia al número total de secuencias en las que se identificaron dos o más PDH. Los patrones que sólo se encuentran en una secuencia portadora son las secuencias "*Singlets*".

Símbolo	Nombre de Patrón	No. Secuencia Portadora
	Cardiobacterium	2
	Anaplasma	1
	Myxococcus	2
	Sulfitobacter	1
	Anaeromyxobacter	14
	Asticcacaulis	1
	Leptothrix	5
	Acidobacterium	19
	Methylobacillus	5
	Holophaga	2
	Methylophaga	6
	Geobacillus	11
	Mariprofundus	15
	Pyramidobacter	2
	Thiothrix	1
	Fusobacterium	1
	Thermovibrio	3
	Thiocystis	5
	Thiorhodovibrio	6
	Meiothermus	9
	Austwickia	11
	Rhodobacteraceae	13
	Beggiatoa	1
	Desulfobacula	3
	Nitrospina	1

Símbolo	Nombre de Patrón	No. Secuencia Portadora
	Thermincola	1
	Desulfomicrobium	1
	Tannerella	24
	Paraprevotella	3
	Caldithrix	1
	Erythrobacter	4
	Ketogulonicigenium	17
	Sphingopyxis	4
	Methylocystis	3
	Thermus	1
	Ralstonia	1
	Diplosphaera	3
	Opitutus	3
	Leadbetterella	2
	Rhodococcus	2
	Bifidobacterium	2
	Plesiocystis	1
	Phycisphaera	1
	Desulfarculus	1
	Conexibacter	2
	Rhodovulum	4
	Rhodomicrobium	3
	Parvularcula	3
	Pedosphaera	2