UNIVERSIDAD NACIONAL AUTÓNOMA DE MEXICO

PROGRAMA DE POSGRADO EN CIENCIAS BIOMÉDICAS

FACULTAD DE MEDICINA

Análisis de la diversidad y evolución de las alcohol deshidrogenasas dependientes de hierro FeADH (tipo III) en eucariontes

Tesis

QUE PARA OPTAR EL GRADO DE DOCTOR EN CIENCIAS

PRESENTA:

CARLOS GAONA LÓPEZ

Director de Tesis:

Dr. Héctor Riveros Rosas

Facultad de Medicina

Ciudad Universitaria, Cd. Mx. **Marzo 2017**

El presente trabajo fue financiado por la Dirección General de Asuntos del Personal Académico, Universidad Nacional Autónoma de México, proyectos PAPIIT IN216513 y 225016 de Dr. Héctor Riveros Rosas, a su vez el Consejo Nacional de Ciencia y Tecnología otorgó una beca de 48 meses a Carlos Gaona López, número de beca 233791, correspondiente a una beca para estudios en el programa de Doctorado en Ciencias Biomédicas, de la Universidad Nacional Autónoma de México.

El presente trabajo fue presentado en dos congresos internacionales, el primero de ellos en el 17th International Workshop on the enzymology and Molecular Biology of Carbonyl Metabolism July 8th to July 13th, 2014 Skytop Lodge Poconos, PA, USA, donde se presentó en la modalidad de poster y short talk, ambas con el nombre de “Structural and phylogenetic analysis support a monophyletic origin of iron–containing alcohol dehydrogenases in animals”

El segundo congreso perteneciente a The Federation of American Societies for Experimental Biology (FASEB), March 28th to April 1th, 2015, Boston Convention and Exhibition Center, Boston Massachusetts, USA, poster “Evolution of the Iron-Containing Alcohol Dehydrogenase in Animals” para este último congreso se contó adicionalmente con recursos del Programa de Apoyo a Estudios de Posgrado, PAEP.

Índice

Índice de figuras y cuadros

Cuadro de abreviaturas

AAD-C subfamilia con doble dominio alcohol y acetaldehído deshidrogenasa.

ADH-Fe alcohol deshidrogenasa dependientes de hierro.

ADH-Zn alcohol deshidrogenasa dependientes de zinc.

ADH- short chain alcohol deshidrogenasa de cadena corta.

Adhfe1 gen de alcohol deshidrogenasa dependiente de hierro humano.

ALDH aldehído deshidrogenasas.

BDH subfamilia butanol deshidrogenasa.

BLAST Basic Local Alignment Search Tool.

DHQ Fe_ADH superfamilia dehidroquinato sintasa alcohol deshidrogenasa dependiente de hierro.

DNA ácido desoxirribonucleico.

DHQ familia dehidroquinato sintasa.

FeADH familia de alcohol deshidrogenasa dependiente de hierro.

GABA ácido γ-aminobutírico.

GHB ácido gamma hidroxibutírico.

GlyDH familia glicerol deshidrogenasa.

G1PDH familia glicerol fosfato deshidrogenasa

HOT subfamilia hidroxiácido oxoácido transhidrogenasa.

$HTU_{(S)}$ unidad taxonómica hipotética.

HVD subfamilia hidroxivalerato deshidrogenasa.

LPO subfamilia lactaldehído propanodiol oxidoreductasa.

MAR subfamilia maleíl acetato reductasa.

MEME modelo de evolución de efectos mezclados.

NADH nicotinamida adenina dinucleótido.

NADPH nicotinamida adenina dinucleótido fosfato.

NADPH-BDH subfamilia butanol deshidrogenasa.

NCBI Centro Nacional para la Información Biotecnológica.

LCA último ancestro común.

$OTU_{(s)}$ unidad taxonómica organizacional.

PDB protein data bank.

PDD subfamilia propanodiol deshidrogenasa.

PGDH prostaglandinas deshidrogenasas.

SSA succinato semialdehído.

UPGMA método de agrupación de pares no ponderado con media aritmética.

VAST vector alignment search tool

Resumen

Las enzimas con actividad alcohol deshidrogenasa se encuentran presentes en organismos de los tres dominios de la vida. Más específicamente, en animales tenemos que dicha actividad se encuentra representada por tres superfamilias de enzimas, las cuales presentan distintos mecanismos de reacción y estructura, sugiriendo que se originaron de manera independiente a lo largo de la historia evolutiva de este linaje. La familia de tipo I o Zn-ADH y la familia de tipo II o ADH de cadena corta, se encuentran bien estudiadas, siendo la familia de tipo III o Fe-ADH de la cual no se han realizado estudios sobre la diversidad y evolución de tal grupo. Por lo tanto, decidimos realizar un estudio, el cual es el primero en ser publicado sobre dicho grupo de enzimas. Nuestro estudio incluye el primer análisis filogenético realizado sobre la familia de Fe-ADH, además de hacer un primer estudio sobre la diversidad de tal familia en eucariontes.

De acuerdo con nuestro análisis filogenético encontramos que las Fe-ADHs pertenecientes a eucariontes pueden agruparse en 13 subfamilias proteicas 8 de las cuales se encuentran representadas en organismos de los tres dominios de la vida, lo que nos indica que se trata de una familia de enzimas muy antigua. Aunque muchas de las enzimas localizadas en este grupo son activadas por átomos de hierro, es normal encontrar una variedad de metales funcionando como cofactores e inclusive la ausencia de cofactor metálico es posible. De las trece subfamilias agrupadas en la familia de FeADH en eucariontes, dos son las más estudiadas y mejor representadas dentro de este dominio de la vida. La primera corresponde a la hidroxiácido oxoácido deshidrogenasa (HOT) presente en todos los animales, algunos grupos de hongos y algunos individuos catalogados en el reino artificial de los protoctistas, además de los dos dominios restantes. Cabe mencionar que los animales presentan enzimas únicamente pertenecientes a la subfamilia HOT, donde juega un papel en la conversión de ácido gamma hidroxibutírico a succinato semialdehído, al menos en mamíferos. También vale la pena mencionar que el gen que codifica para dicha enzima es de copia única, además dicha proteína tiene una ubicación celular, en la mitocondria, teniendo como respaldo evidencia tanto experimental como aquella proporcionada por distintos softwares de predicción. En cuanto a la segunda subfamilia, las maleíl-acetato reductasas (MAR), corresponde a enzimas pertenecientes principalmente a organismos del reino Fungi donde cumplen la importante función de degradar aquellos metabolitos del rompimiento del anillo “ring-fission” que derivan de la degradación aeróbica de compuestos aromáticos en microorganismos. Por último destacamos que organismos pertenecientes al reino Plantae presentan enzimas agrupadas en 3 distintas subfamilias proteicas, dichas enzimas se encuentran restringidas al grupo de las clorofitas, dejando a las plantas superiores ausentes de cualquier enzima perteneciente a las FeADHs.

En conclusión las FeADHs son una antigua y diversa familia de enzimas que comparten un plegamiento tridimensional común, hecho que fue aprovechado por nosotros para realizar un alineamiento estructural que sirvió como base de nuestro análisis filogenético. Un aspecto curioso es la distribución irregular de las FeADHs en eucariontes, donde prácticamente el 85% de las secuencias se agrupan en las subfamilias HOT y MAR.

Abstract

Background

Alcohol dehydrogenase (ADH) activity is widely distributed in the three domains of life. Currently, there are three non-homologous $NAD(P)^+$-dependent ADH families reported: Type I ADH comprises Zn-dependent ADHs; type II ADH comprises short-chain ADHs described first in Drosophila; and, type III ADH comprises iron-containing ADHs (FeADHs). These three families arose independently throughout evolution and possess different structures and mechanisms of reaction. While types I and II ADHs have been extensively studied, analyses about the evolution and diversity of (type III) FeADHs have not been published yet. Therefore in this work, a phylogenetic analysis of FeADHs was performed to get insights into the evolution of this protein family, as well as explore the diversity of FeADHs in eukaryotes.

Principal Findings

Results showed that FeADHs from eukaryotes are distributed in thirteen protein subfamilies, eight of them possessing protein sequences distributed in the three domains of life. Interestingly, none of these protein subfamilies possess protein sequences found simultaneously in animals, plants and fungi. Many FeADHs are activated by or contain Fe^{2+}, but many others bind to a variety of metals, or even lack of metal cofactor. Animal FeADHs are found in just one protein subfamily, the hydroxyacid-oxoacid transhydrogenase (HOT) subfamily, which includes protein sequences widely distributed in fungi, but not in plants), and in several taxa from lower eukaryotes, bacteria and archaea. Fungi FeADHs are found mainly in two subfamilies: HOT and maleylacetate reductase (MAR), but some can be found also in other three different protein subfamilies. Plant FeADHs are found only in chlorophyta but not in higher plants, and are distributed in three different protein subfamilies.

Conclusions/Significance

FeADHs are a diverse and ancient protein family that shares a common 3D scaffold with a patchy distribution in eukaryotes. The majority of sequenced FeADHs from eukaryotes are distributed in just two subfamilies, HOT and MAR (found mainly in animals and fungi). These two subfamilies comprise almost 85% of all sequenced FeADHs in eukaryotes.

1.1 Bases de datos biológicas

Las bases de datos biológicas son recopilaciones de información sobre aspectos biológicos. Dicha información es recogida de experimentos científicos, literatura publicada, y análisis computacionales (Attwood *et al.,* 2011). Contienen información de áreas de investigación incluyendo genómica, proteómica, metabolómica, expresión génica mediante microarreglos, y filogenética. (Altman 2004; Bourne 2005). La información contenida en dichas bases de datos incluye aspectos tales como funciones, estructura y localización (tanto celular como cromosómica) de genes, así como similitudes de secuencias y estructuras biológicas (Attwood *et al.,* 2011; Bourne, 2005).

Podemos clasificar a las bases de datos biológicas ampliamente como bases de datos de secuencias y bases de datos de estructura (Attwood *et al.,* 2011; Bourne, 2005).

Las secuencias de ácidos nucleicos y proteínas son almacenadas en bases de datos de secuencias mientras que las bases de datos de estructuras almacenan principalmente información de proteínas, aunque el número de estructuras de ácidos nucleicos ha crecido en años recientes. Tales bases de datos son útiles para comprender y explicar una serie de fenómenos biológicos que van desde la estructura de proteínas y su interacción, hasta el metabolismo completo de los organismos, además, de ser útiles en la comprensión de la evolución de las especies (Bourne, 2005; Zou *et al.,* 2015).

Otra manera de clasificar a las bases de datos es de acuerdo a la fuente de los datos que estas contienen, pudiendo encontrar bases de datos primarias y bases de datos secundarias; estas últimas, contienen datos e hipótesis derivados del análisis de las bases de datos primarias, como son mutaciones, relaciones evolutivas, agrupación por familias o funciones, entre otras (Bourne, 2005).

- Las bases de datos primarias; contienen información directa de la secuencia, estructura o patrón de expresión ya sea de un gen o una proteína, es decir que

los datos experimentales son directamente subidos a las bases de datos. En esta categoría encontramos las bases de datos GenBank, DNA Data Bank of Japan (DDJB), UniProtKB/SwissProt, UniProtKB/TrEMBL y Protein Data Bank (PDB) (Altman, 2004).

- Las bases de datos secundarias; contienen información derivada de las bases de datos primarias. Una base de datos secundaria de secuencias contiene información de la conservación de la secuencia, patrones de secuencia y residuos del sitio activo de familias de proteínas derivados de alineamientos múltiples entre secuencias evolutivamente relacionadas. Una base de datos secundaria de estructuras organiza las entradas del PDB clasificándolas, por ejemplo, de acuerdo a su estructura. Algunos ejemplos de éstas bases de datos son: CATH y SCOP (Altman, 2004; Andreeva *et al.,* 2008).

En cuanto al origen de las bases de datos biológicas, muchos autores concuerdan en que, dado el perfeccionamiento y desarrollo de técnicas experimentales de alto rendimiento como son la secuenciación del DNA, la cristalografía de rayos X, así como los análisis de microarreglos, por citar las principales; generó un crecimiento exponencial en la cantidad de datos biológicos obtenidos (secuencias genómicas y de proteínas, estructuras de proteínas, expresión génica, etc.) que generaron la necesidad de contar con formas eficientes de almacenar la información (Zou *et al.,* 2015).

Un problema fundamental en todas las bases de datos de secuencias, es que los registros provienen de una gran variedad de fuentes. Como resultado, las secuencias mismas y principalmente las anotaciones biológicas adjuntas a estas secuencias, varían notablemente en calidad. Además, tienen mucha redundancia ya que muchos laboratorios ingresan a menudo secuencias que son idénticas o muy similares a otras en las bases de datos (Attwood *et al.,* 2011; Bolser *et al.,* 2012).

Un recurso importante para la búsqueda de bases de datos biológicos es la edición anual de la revista Nucleic Acids Research (NAR). Ésta edición de bases

de datos en NAR está disponible gratuitamente todos los años, donde se publican nuevas bases de datos y algunas actualizaciones de las ya conocidas, figura 1. Se encuentran clasificadas de acuerdo a su temática y están en línea a disposición de toda la comunidad científica (Zou *et al.,* 2015; Brazas *et al.,* 2011).

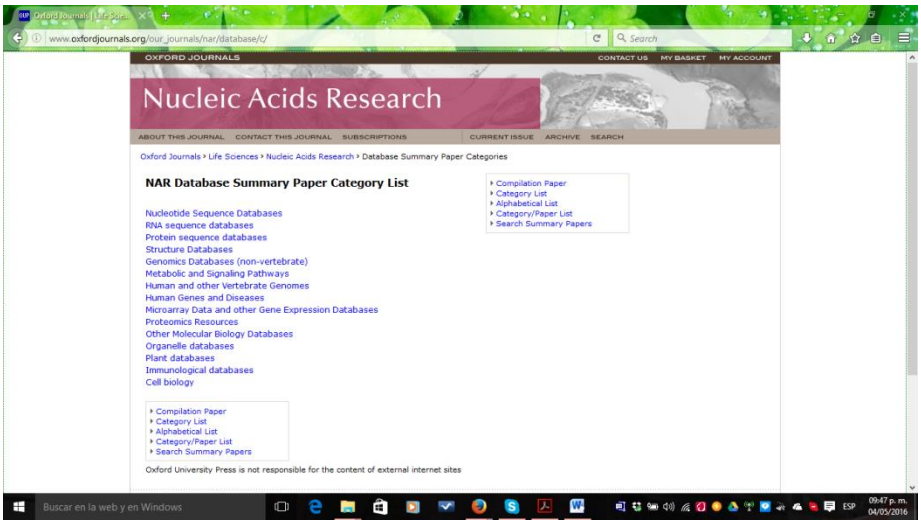


Figura 1. Captura de pantalla de la página web de la revista Nucleic Acids Research, https://www.oxfordjournals.org/our_journals/nar/database/c/

A continuación se presentan algunas de las principales bases de datos biológicas disponibles.

Cuadro 1. Bases de datos de secuencias proteicas (Bolser *et al.,* 2012).

UniProt **Uni**versal **Prot**ein resource (EBI, Swiss Institute of Bioinformatics, PIR) **http://www.uniprot.org/**
Protein Information Resource (Georgetown University Medical Center (GUMC)) **http://pir.georgetown.edu/**
Swiss-Prot Protein Knowledgebase (Swiss Institute of Bioinformatics) **http://www.sib.swiss/**
PEDANT **P**rotein **E**xtraction, **D**escription and **AN**alysis **T**ool (Forschungszentrum f. Umwelt & Gesundheit) **http://pedant.gsf.de/**
PROSITE Database of Protein Families and Domains **http://prosite.expasy.org/**
Database of Interacting Proteins (Univ. of California) **http://dip.doe-mbi.ucla.edu/dip/Main.cgi**
Pfam **P**rotein **fam**ilies database of alignments and HMMs (Sanger Institute)

http://pfam.xfam.org/
PRINTS a compendium of protein fingerprints from (Manchester University) **http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/index.php**
ProDom Comprehensive set of **Pro**tein **Dom**ain Families (INRA/CNRS) **http://prodom.prabi.fr/prodom/current/html/home.php**
SignalP 4.1 Server for signal peptide prediction (including cleavage site prediction), based on artificial neural networks and HMMs **http://www.cbs.dtu.dk/services/SignalP/**
SUPERFAMILY Library of HMMs representing superfamilies and database of (superfamily and family) annotations for all completely sequenced organisms **http://supfam.org/SUPERFAMILY/**
neXtProt - a human protein centric knowledge resource **https://www.nextprot.org/**
Annotation Clearing House a project from the National Microbial Pathogen Data Resource **http://www.nmpdr.org/FIG/wiki/view.cgi**
InterPro Classifies proteins into families and predicts the presence of domains and sites. **https://www.ebi.ac.uk/interpro/**
ProteomeScout - Includes a graphics exports of protein annotations including domains, secondary structure, and post-translational modifications **https://proteomescout.wustl.edu/**

Cuadro 2. Bases de datos de modelos proteicos (Bolser *et al.*, 2012).

Swiss-model Server and Repository for Protein Structure Models **https://swissmodel.expasy.org/repository**
ModBase Database of Comparative Protein Structure Models (Sali Lab, UCSF) **https://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi**
Protein Model Portal (PMP) Meta database that combines several databases of protein structure models (Biozentrum, Basel, Switzerland) **http://www.proteinmodelportal.org/**
Similarity Matrix of Proteins (SIMAP) is a database of protein similarities computed using FASTA. **http://cube.univie.ac.at/resources/simap**

Cuadro 3. Bases de datos de estructuras proteicas (Bolser *et al.,* 2012).

Protein Data Bank (PDB) comprende:

Protein DataBank in Europe (PDBe) **http://www.ebi.ac.uk/pdbe/docs_dev/nobel2012/**
Protein Databank in Japan (PDBj) **https://pdbj.org/**
Research Collaboratory for Structural Bioinformatics (RCSB) **http://www.rcsb.org/pdb/home/home.do**

Cuadro 4. Bases de datos secundarias otras (Bourne 2005; Zou *et al.,* 2015).

SCOP Structural Classification of Proteins **http://scop.mrc-lmb.cam.ac.uk/scop/**
CATH Protein Structure Classification **https://www.ucl.ac.uk/orengo-group/resources/cath-protein-structure-classification**
PDBsum **http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=index.html**

Cuadro 5. Bases de datos de función de proteínas y rutas metabólicas (Bolser *et al.,* 2012).

BioCyc Database Collection including EcoCyc and MetaCyc **https://biocyc.org/**
BRENDA The Comprehensive Enzyme Information System, including FRENDA, AMENDA, DRENDA, and KENDA. **http://www.brenda-enzymes.org/**
KEGG PATHWAY Database (Univ. of Kyoto) **http://www.genome.jp/kegg/pathway.html**
MANET database (University of Illinois) **http://manet.illinois.edu/**
MetaboLights Metabolomics experiments and derived information: metabolite structures, reference spectra, biological roles, locations and concentrations. (European Bioinformatics Institute)

http://www.ebi.ac.uk/metabolights/
MetaNetX Automated Model Construction and Genome Annotation for Large-Scale Metabolic Networks **http://www.metanetx.org/**
Reactome Navigable map of human biological pathways, ranging from metabolic processes to hormonal signalling. (Cold Spring Harbor Laboratory, European Bioinformatics Institute, Gene Ontology Consortium) **http://www.reactome.org/**
Small Molecule Pathway Database (SMPDB) **http://smpdb.ca/**

Cuadro 6. Bases de datos de microarreglos (Bolser *et al.*, 2012).

ArrayExpress (European Bioinformatics Institute) **https://www.ebi.ac.uk/arrayexpress/**
Gene Expression Omnibus (GEO, National Center for Biotechnology Information) **https://www.ncbi.nlm.nih.gov/geo/**
GPX(Scottish Centre for Genomic Technology and Informatics) **http://gpxmea.gti.ed.ac.uk/**
maxd (Univ. of Manchester) **http://bioinf.man.ac.uk/microarray/maxd/**
Stanford Microarray Database (SMD) (Stanford University) **http://www.http.com//genome-www.stanford.edu/microarray**
Genevestigator - Expression Search Engine (Nebion AG) **https://genevestigator.com/gv/**
Bgee Bgee is a database to retrieve and compare gene expression patterns between species. It only contains wild-type and manually curated microarray/RNASeq/in situ experiments. **https://bgeedb.wordpress.com/**
BioGPS (The Scripps Research Institute) A Gene Portal System with a Gene Expression Visualizer. **http://biogps.org/#goto=welcome**
The European Genome-phenome Archive (EGA) **https://www.ebi.ac.uk/ega/home**

1.2 Alineamiento múltiple de secuencias

Quizá la mayor contribución de las secuencias biológicas a los análisis evolutivos está dada por el hecho de que secuencias de diferentes organismos están a menudo relacionadas. Por ejemplo, genes o proteínas similares están conservados entre especies ampliamente divergentes, muchas veces realizando una función similar o incluso idéntica, y otras veces, mutando o rearreglándose para realizar una función alterada por efecto de la fuerza de selección natural. Por lo tanto, muchos genes están representados de una manera altamente conservada en una amplia gama de organismos. Es por eso que mediante un alineamiento múltiple de secuencias de estos genes y/o secuencias proteicas, podemos analizar los patrones de cambio que han ocurrido a lo largo de la historia evolutiva, obteniendo además, información importante como la ancestría común entre dos o más genes o secuencias. (Mount, 2004).

Por definición, un alineamiento múltiple de secuencias es un alineamiento de tres o más secuencias biológicas, ya sea de ácidos nucleicos (DNA o RNA) o proteínas. En la mayoría de los casos, el conjunto de entrada de secuencias que se desea consultar es denominado como secuencias problema, y se asume que tiene una relación evolutiva, diciéndose que ellas comparten un linaje por lo que descienden de un ancestro común. De un alineamiento múltiple de secuencias, se puede inferir por lo tanto homología de secuencia y por ende, puede llevarse a cabo un análisis filogenético para evaluar el origen evolutivo compartido entre ellas. En dichos análisis se pueden observar eventos como mutaciones puntuales, denotado por cambios en un solo nucleótido o aminoácido, así como indels (inserciones o deleciones). Un alineamiento múltiple de secuencias puede ser usado para evaluar la conservación en las secuencias de dominios proteicos, así como de estructuras secundarias y terciarias, e incluso de aminoácidos individuales o nucleótidos (Mount, 2004; Wallace, 2005; Bailey, 1994; Letunic *et al.,* 2012).

Posteriormente a partir de un alineamiento múltiple de secuencias podemos obtener un árbol filogenético, esto debido a dos razones. La primera es porque los dominios funcionales que son conocidos en secuencias anotadas se pueden utilizar para la alineación en secuencias problema. La segunda es que dichos análisis pueden ser usados para identificar sitios funcionalmente importantes, como son; sitios de unión, sitios activos o algún otro sitio correspondiente a una función clave mediante la localización de dominios conservados (Costantini *et al.,* 2006; Bailey, 1994; Budd, 2009).

Al analizar alineamientos múltiples, es útil tener en cuenta los diferentes aspectos de las secuencias al compararlas entre ellas. Estos aspectos incluyen identidad, similitud y homología. “Identidad” significa que las secuencias tienen residuos idénticos en sus respectivas posiciones. Mientras que “similitud” tiene que ver con residuos cualitativamente similares en sus respectivas posiciones al comparar las secuencias. Por ejemplo, al hablar de secuencias de nucleótidos, las pirimidinas son consideradas similares entre ellas, de igual manera pasa con las purinas. Posteriormente dicha similitud conduce a la homología, ya que entre mayor similitud exista entre un número de secuencias dado, más cerca están éstas de ser homólogas. Por lo que la similitud de secuencias es el principal criterio para inferir una ancestría común entre un grupo de secuencias (Budd, 2009; Bailey, 1998).

1.3 Alineamiento estructural

Un alineamiento estructural es un tipo de alineamiento de secuencias basado en la comparación de la forma de dos o más estructuras Estos alineamientos intentan establecer equivalencias entre dos o más estructuras generalmente, proteínas basándose en conformación tridimensional. El proceso se aplica normalmente a las estructuras terciarias de las proteínas, pero también puede usarse para moléculas de RNA. En contraste a la simple superposición estructural, donde al menos se conocen algunos residuos equivalentes de las dos estructuras, el alineamiento estructural no requiere un conocimiento previo de posiciones

equivalentes. Suele usarse un alineamiento estructural cuando se comparan proteínas con baja similitud entre sus secuencias, en donde las relaciones evolutivas entre proteínas no pueden ser tan fácilmente detectada por técnicas estándares de alineamiento de secuencias.

El alineamiento estructural puede usarse, por lo tanto, para sugerir relaciones evolutivas entre proteínas que comparten una secuencia con muy baja similitud. Sin embargo, el uso de estos resultados como evidencia de un ancestro evolutivo común debe realizarse con cautela debido a la posibilidad de evolución convergente, según la cual múltiples secuencias de aminoácidos sin relación filogenética entre si convergen a una misma estructura terciaria (Zhang *et al.,* 2018; Zemla, 2003).

Los alineamientos estructurales pueden comparar dos o múltiples secuencias. Dado que estos alineamientos dependen de información sobre todas las conformaciones tridimensionales de las secuencias problema, el método sólo puede ser usado sobre secuencias donde estas estructuras sean conocidas. Estas se determinan normalmente por cristalografía de rayos X o espectroscopia de resonancia magnética nuclear. Por otro lado, también es posible realizar un alineamiento estructural sobre estructuras obtenidas mediante métodos de predicción de estructura. Los alineamientos estructurales son especialmente útiles para analizar datos surgidos de los campos de la genómica estructural y de la proteómica, y pueden usarse como puntos de comparación para evaluar alineamientos generados por métodos bioinformáticas basados exclusivamente en secuencias (Wallace *et al*., 2005; Zhang *et al*., 2005).

El resultado de un alineamiento estructural es una superposición de los conjuntos de coordenadas atómicas, así como una distancia media cuadrática mínima (o RMSD, de Root Mean Square Deviation, o desviación de la media cuadrática) entre las estructuras básicas de las proteínas superpuestas. La RMSD de estructuras alineadas indica la divergencia entre ellas. El alineamiento estructural presenta problemas con la existencia de múltiples dominios proteicos en el interior de una o más de las estructuras de entrada, ya que cambios en la orientación

relativa de los dominios entre dos estructuras a alinear pueden exagerar la RMSD artificialmente (Zemla, 2003; Taylor *et al.,* 1994).

Como ya se mencionó, la información producida por un alineamiento estructural es un conjunto de coordenadas tridimensionales superpuestas para cada estructura a comparar. Normalmente uno de los elementos de entrada puede estar fijado como referencia y que, por lo tanto, sus coordenadas superpuestas no cambiaran. Las estructuras encajadas pueden usarse para calcular valores RMSD mutuos, así como otras medidas de similitud estructural más sofisticadas como el test de distancia global (GDT, de sus siglas en inglés, y que es la prueba utilizada en CASP, Critical Assessment of Techniques for Protein Structure Prediction). Un alineamiento estructural también implica un alineamiento de secuencias unidimensional desde el que una secuencia identidad, o el porcentaje de residuos que son idénticos entre las estructuras de entrada, puede calcularse como una medida de cuan cercanamente se encuentran ambas secuencias (Zhang *et al*., 2008; Zemla, 2003).

Cuando se alinean estructuras con secuencias muy diferentes, los átomos de la cadena lateral, generalmente, no se toman en cuenta, ya que sus identidades difieren en muchos de los residuos alineados. Por esta razón, en los métodos de alineamiento estructural es común usar por defecto solo los átomos del esqueleto incluidos en el enlace peptídico. Por simplicidad y eficiencia a menudo solo se consideran las posiciones del carbono alfa, ya que el enlace peptídico tiene una conformación planar con muy poca variación. Solo cuando las estructuras a alinear son altamente similares, e incluso idénticas, es significativo alinear posiciones de átomos de la cadena lateral, en cuyo caso la RMSD refleja no solo la conformación del esqueleto de la proteína, sino también los estados de las rotaciones angulares en las cadenas laterales. Otros criterios de comparación que reducen el ruido e impulsan las coincidencias incluyen tomar en consideración las estructuras secundarias de las proteínas (Zemla, 2003; Taylor *et al.,* 1994; Godzik, 1996).

La comparación más sencilla posible entre estructuras comprende a la superposición estructural, la cual es usada para comparar conformaciones múltiples de la misma proteína (en cuyo caso no es necesario el alineamiento ya que la secuencia es la misma) y para evaluar la calidad de los alineamientos producidos usando sólo información de las secuencias entre dos o más secuencias cuyas estructuras son conocidas. Este método utiliza tradicionalmente un sencillo algoritmo de ajuste por mínimos cuadrados, en el que las rotaciones y translaciones óptimas se encuentran minimizando la suma de los cuadrados de las distancias entre todas las estructuras de la superposición. Más recientemente, los métodos bayesianos y de máxima verosimilitud han incrementado enormemente la precisión de las rotaciones, translaciones y matrices de covarianza estimadas para la superposición (Godzik, 1996; Wang *et al*., 1994).

Cuadro 7. Se presentan algunos softwares para realizar alineamiento estructural.

Software para alineamiento estructural	Descripción	Clase
VAST Vector Alignment Search Tool	Alineamiento de elementos de estructura secundaria	Alineamiento de pares https://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml
MatAlign Matrix Alignment	Comparación de estructuras de proteínas por Matrix Alignment (Mapa de contacto)	Alineamiento de pares http://stormo.wustl.edu/MatAlign/
MAMMOTH MAtching Molecular Models Obtained from Theory	Alineamiento por átomo (Cα) de esqueleto	Alineamiento de pares http://physbio.mssm.edu/~ortizg/
MASS Multiple Alignment by Secondary Structure	Alineamiento de elementos de estructura secundaria	Alineamiento múltiple de estructuras http://bioinfo3d.cs.tau.ac.il/MASS
SCALI Structural Core ALIgnment of proteins	Alineamiento basado en secuencia	Alineamiento de pares http://www.bioinfo.rpi.edu/bystrc/scali.html
CAALIGN	Alineamiento por átomo (Cα) de esqueleto	Alineamiento múltiple de estructuras http://www.ebi.ac.uk/node/13123

1.4 Predicción de la estructura de proteínas

Una de las principales metas del análisis de secuencias proteicas es comprender la relación entre la secuencia de aminoácidos y la estructura tridimensional de una proteína. Si esta relación fuera conocida, la estructura de una proteína podría ser seguramente predicha a partir de la secuencia de aminoácidos. Por desgracia, la relación existente entre secuencia y estructura no es tan sencilla de evaluar. Se ha

avanzado mucho en la categorización de proteínas sobre la base de la estructura o secuencia, y este tipo de información es muy útil para el modelado de proteínas. Como veremos más adelante, esta información se ha usado para el desarrollo de programas de cómputo que buscan dilucidar la estructura terciaria de una proteína a partir de su secuencia de aminoácidos (Mount, 2004).

Actualmente, la brecha entre secuencias proteicas conocidas y la estructura proteica de tales secuencias es inmensa. A comienzos del 2008, solamente cerca del 1% de las secuencias listadas en la base de datos del UniProtKB correspondían a estructuras en el Protein Data Bank (PDB), dejando un hueco entre las secuencias conocidas y las estructuras resueltas de casi 5 millones de secuencias de las cuales no se ha resuelto su estructura tridimensional (Rigden, 2009). Al día de elaboración de esta tesis se encontraron registradas en la página oficial del PDB (http://www.rcsb.org/pdb/home/home.do) 125,093 estructuras macromoleculares de las cuales 116,105 pertenecían a proteínas.

Las distintas técnicas experimentales para determinar la estructura terciaria de una proteína han encarado serias dificultades en su habilidad para determinar estructuras de proteínas particulares. Por ejemplo, mientras que la cristalografía de rayos X ha sido bastante exitosa en dilucidar la estructura de aproximadamente 80 000 proteínas citosólicas, esta misma ha tenido mucho menos éxito en la cristalización de proteínas de membrana con aproximadamente 280 estructuras solamente. Dadas estas limitaciones experimentales el desarrollo de programas informáticos para acortar la brecha entre las secuencias y estructuras proteicas conocidas se ve como uno de los campos más promisorios a desarrollar en el futuro cercano (Yonath, 2011).

La predicción de la estructura tridimensional de una proteína a partir de su secuencia de aminoácidos, es decir la predicción de su plegamiento así como su estructura secundaria, terciaria e incluso cuaternaria a partir de su estructura primaria, es una de las más importantes metas perseguida por la bioinformática (Zhang *et al*., 2005; Yonath, 2011).

Actualmente existen dos estrategias básicas para aproximarse a la predicción de la estructura de una proteína: la predicción *de novo* también conocida como predicción *ab initio*, en la que se suelen utilizar métodos estocásticos; y la predicción por comparación, en la que se recurre a una biblioteca de estructuras previamente conocidas; cabe mencionar que dichas estructuras han sido resueltas generalmente por técnicas de cristalografía de rayos X y espectroscopía de resonancia magnética nuclear (Zhang *et al.,* 2005).

La predicción *ab initio* se basa en el hecho de que el plegamiento de una proteína no puede ser aleatorio; por ejemplo en una proteína de sólo 100 aminoácidos donde cada aminoácido puede tomar 10 diferentes conformaciones en promedio, tenemos que puede haber 10^{100} diferentes conformaciones para un polipéptido relativamente pequeño como éste. Si cada posible conformación fuera probada cada 100 picosegundos (10^{-13}), esto le tomaría a una proteína cerca de 10^{77} años probar todas las conformaciones posibles siendo, más que obvio, incompatible con la vida. Los métodos de predicción de estructura proteica *de novo* intentan predecir la estructura terciaria de una secuencia basados en los principios generales que gobiernan la energética del plegamiento de proteínas así como las tendencias estadísticas de las características conformacionales que las estructuras nativas adquieren, sin el uso de templados explícitos (Guzzo, 1965; McDonald *et al*., 2001).

Por otro lado, los modelos de predicción por comparación usan estructuras resueltas previamente, como punto de partida o templados. A pesar de la vasta cantidad de proteínas descritas, casi todas pueden ser agrupadas en un limitado juego de motivos estructurales terciarios. Inclusive se ha sugerido que hay solamente alrededor de 2000 distintos plegamientos proteicos en la naturaleza, en los cuales pueden ser clasificados los varios millones de proteínas (Zhang *et al.,* 2005; Dor y Zhou, 2006).

A su vez estos métodos pueden también ser divididos en dos categorías:

Modelado por homología

Dicho método se basa en el principio de que sí dos proteínas son homologas deberían presentar estructuras muy similares. Esto debido a que el plegamiento de una proteína está mejor conservado evolutivamente que su secuencia de aminoácidos, por lo que una secuencia problema puede ser modelada con razonable precisión tomando como base templados, aún de proteínas distantemente relacionadas. Siempre y cuando la relación entre ambas secuencias, problema y templado, pueda ser discernida a través de un alineamiento de secuencias, ya que como se mencionó anteriormente está el problema de convergencia evolutiva, donde dos proteínas sin relación aparente presentan una muy parecida estructura. De hecho, según los expertos, se ha propuesto que el principal cuello de botella en los modelados comparativos surge de dificultades en el alineamiento más que de errores en la predicción de una estructura. No es de sorprender que los modelados por homología sean más precisos cuando las secuencias problema y templado presentan mayor similitud o identidad (Zhang *et al.,* 2005).

Enhebrado de proteínas (Protein threading)

También es conocido como método de reconocimiento de plegamiento, dicho método es usado para modelar aquellas secuencias proteicas problema las cuales tienen el mismo plegamiento que proteínas de estructura conocida, pero no se cuenta con estructuras resueltas de proteínas homologas a la secuencia problema. El enhebrado funciona mediante el uso de modelos estadísticos de la relación de las estructuras depositadas en el PDB y la secuencia de la proteína que uno desea modelar (Bowie *et al.,* 1991; Dunbrack, 2006; Brylinski, 2013).

Tal predicción es hecha por enhebrado, es decir localización y alineamiento, donde cada aminoácido de una secuencia problema es probado en la estructura templado, y evaluando cuan bien encaja en dicha posición sobre el templado. Después de que el templado que mejor encaja es seleccionado, el modelo

estructural de la secuencia problema es construido tomando como base el alineamiento con el o los templados elegidos. Dicho método tiene como respaldo dos aspectos muy importantes; el primero es que el número de plegamientos diferentes en la naturaleza es bastante pequeño, entre 1,300 y 2,000 plegamientos distintos; segundo que casi el 90% de las nuevas estructuras subidas al PDB en los últimos años presentan plegamientos similares a las previamente depositadas en el PDB (Bowie *et al.,* 1991; Jones *et al*., 1992; Dunbrack, 2006; Brylinski, 2013).

1.5 Filogenética y arboles filogenéticos

Para Lineo y sus sucesores, el objetivo último de la sistemática y de la practica taxonómica consistía en describir el plan o las ideas con las que Dios había construido el mundo natural. Por ejemplo, una única vertebra repetida en número y formas variables era el plan básico que el Creador había formulado para la construcción de todos los animales cordados. No fue sino hasta después de la publicación de “Sobre el Origen de las Especies” cuando las similitudes y las diferencias entre los organismos fueron interpretadas desde una perspectiva biológica (Strickberger y Monroe, 1996).

Las similitudes entre organismos podían ser agrupadas en dos tipos: ya sea que fuesen procesos de adaptación independiente a ambientes semejantes, o eran la evidencia de que ambos organismos habían tenido un pasado en común. Charles Darwin hacía referencia a estas similitudes como debidas a las “condiciones de existencia” o como consecuencia de una “unidad tipo”. Hoy nos referimos a estas similitudes como analogías y homologías, respectivamente, y las diferencias entre ambas es un concepto clave para la formación de grupos inclusivos (Strickberger y Monroe, 1996).

El concepto de filogenia, por lo tanto, hace referencia a la historia evolutiva de las especies. Un análisis filogenético es un medio para estimar las relaciones evolutivas, en los análisis de filogenética molecular, la secuencia de un gen o proteína puede ser usada para examinar las relaciones evolutivas entre diferentes especies. Las relaciones evolutivas obtenidas de análisis filogenéticos se

representan usualmente como ramificaciones de un diagrama en forma de árbol, conocido como árbol filogenético (Choudhuri, 2014).

1.5.1 Árboles filogenéticos

La filogenia de cualquier grupo de especies puede ser representada por un árbol filogenético (Figura 2), dicho árbol no es más que un diagrama con ramificaciones compuesto de nodos y ramas. Este diagrama representa una hipótesis de las relaciones de ancestría descendencia de los taxa que contiene. Cada árbol presenta una topología característica, denotado por su patrón de ramificación.

Los nodos, en un árbol filogenético, representan unidades taxonómicas, tal como especies o taxa superiores, poblaciones, genes o proteínas. Un rasgo importante en un árbol filogenético es la longitud de cada una de sus ramas, las cuales representan el tiempo estimado de divergencia entre las unidades taxonómicas. Una rama puede conectar solamente dos nodos. En dicho diagrama, los nodos terminales representan lo que se conoce como unidad taxonómica operacional (OTUs) u hojas. Cabe mencionar que tales OTUs representan el objeto actual; ya sea especies, poblaciones e inclusive secuencias génicas o proteicas, que se desean comparar, por lo que los nodos internos representan unidades taxonómicas hipotéticas (HTUs). Una HTU es una unidad inferida la cual representa el último ancestro común (LCA) a partir del cual ambos nodos surgieron. Aquellos (OTUs) que divergieron de un solo nodo forman grupos hermanos, mientras que un taxón que cae fuera del clado se le conoce como grupo externo. Por ejemplo, en la figura 2B, T_2 y T_3 son grupos hermanos, mientras que T_1 es un grupo externo de T_2 y T_3 (Choudhuri, 2014).

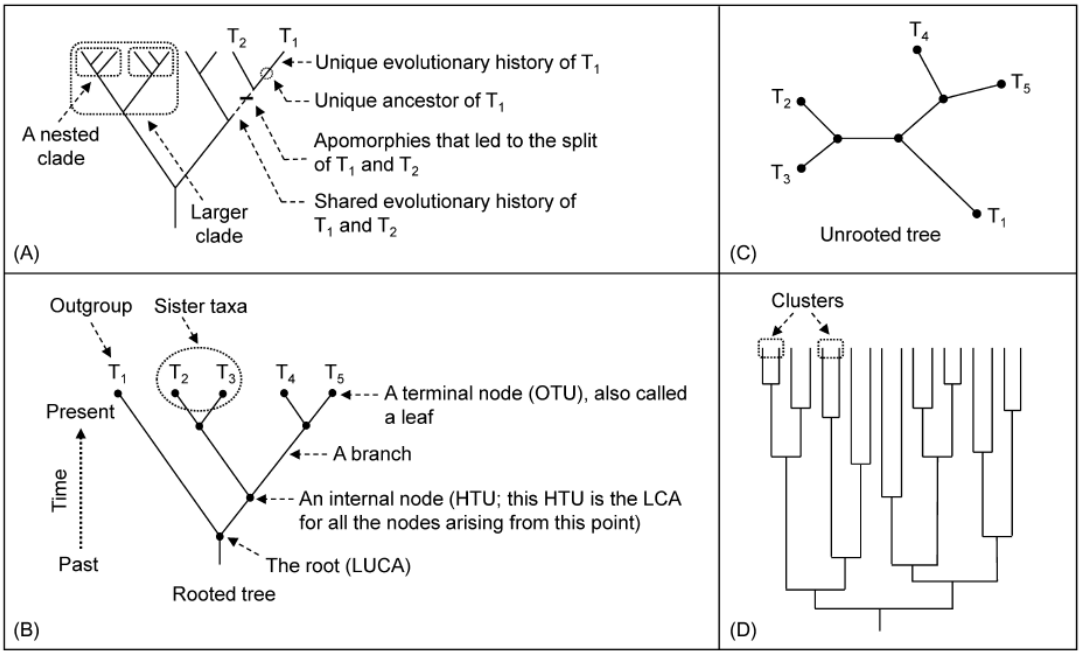


Figura 2. Diferentes formas de representar un árbol filogenético. En el cuadro D se ejemplifica un dendrograma derivado de un agrupamiento jerárquico. A, B y D son ejemplos de árboles enraizados, mientras que C muestra un árbol sin enraizar. Los taxa que comparten caracteres derivados específicos son agrupados dentro de clados. (A) clados pequeños localizados dentro de un clado mayor son conocidos como clados anidados. (B) los nodos terminales representan la unidad taxonómica operacional, también llamada hojas; cada nodo terminal podría ser un taxón (especies o taxa superior) o bien una secuencia génica o proteica. Los nodos internos representan unidades taxonómicas hipotéticas, por sus siglas en inglés HTU. Un HTU representa el último ancestro común para todos los nodos que surgieron de ese punto. Mientras que dos descendientes que se separan del mismo nodo son llamados grupos hermanos y un taxón que cae fuera del clado es llamado grupo externo. Los árboles enraizados tienen un nodo del cual el resto del árbol diverge, frecuentemente llamado el último ancestro común o por sus siglas en inglés LUCA. Imagen tomada de BIOINFORMATICS FOR BEGINNERS de Choudhuri, 2014.

1.5.2 Principios básicos en la construcción de un árbol filogenético

Hoy en día existen una gran variedad de recursos en línea para aquellos que desean construir un árbol filogenético, estos recursos están dirigidos desde aquellos quienes realizan por primera un árbol filogenético hasta gente experta en el tema. Independientemente del grado de práctica que uno posea, la construcción de árboles filogenéticos implica ciertas suposiciones y pasos.

Existen ciertos supuestos a la hora de construir un árbol filogenético, tal como; a) la suposición de homología entre las secuencias de estudio, b) la suposición de

que cada secuencia evolucionó independientemente de las demás, c) por último, es importante asumir la ausencia de transferencia horizontal de genes. De acuerdo a los expertos un paso crucial en la obtención de cualquier árbol filogenético idóneo, es tener un alineamiento múltiple confiable, a partir del cual se obtiene dicho árbol. Cuando se usan secuencias codificantes, es preferible usar las secuencias proteicas para reconstruir el árbol filogenético (Mount, 2004; Hall, 2011).

La construcción de un árbol filogenético involucra los siguientes pasos: 1) Selección del marcador molecular apropiado (genes/proteínas/DNA mitocondrial), 2) Un alineamiento de secuencias múltiple, 3) Selección de un modelo de Evolución, 4) Construcción de un árbol filogenético, 5) Evaluación de la fiabilidad del árbol (Mount, 2004; Lemey *et al*., 2009; Fitch *et al.,* 1967; Hall, 2011).

1.5.2.1 Selección del marcador molecular apropiado

La elección de secuencias de proteínas o ácidos nucleicos como marcador depende mucho de las necesidades que se tengan. Un marcador molecular en un análisis filogenético es la información biológica que es usada para inferir relaciones evolutivas entre taxa (Hall, 2011). Cuando se trata de secuencias codificantes, los expertos sugieren usar las secuencias proteicas, como marcador molecular para la construcción de un árbol filogenético. Algunas de las razones del porque las secuencias proteicas son más apropiadas sobre las secuencias de DNA son las siguientes:

1. Existen un mayor número de estados de carácter para los distintos aminoácidos (20) que para los nucleótidos (4).

2. Las matrices de sustitución de aminoácidos son mucho más sofisticadas que las matrices de sustitución de nucleótidos (Mount, 2004; Lemey *et al.,* 2009; Hall, 2011).

3. La existencia de codones preferenciales para el mismo aminoácido en diferentes especies puede artificialmente sobreestimar la variación en las secuencias de nucleótidos (Mount, 2004; Hall, 2011).

Pese a las desventajas que han sido citadas previamente, las secuencias de nucleótidos pueden ser usadas también como marcadores moleculares. De hecho, es recomendable utilizar secuencias de nucleótidos, las cuales evolucionan más rápidamente que las proteínas, cuando se estudian organismos estrechamente relacionados por ejemplo, usando regiones no codificantes de DNA mitocondrial. Otro ejemplo es cuando se comparan proteínas cuyas secuencias están altamente conservadas entre especies, en este caso el principio es que, dadas dos secuencias proteicas que son idénticas, su secuencia de nucleótidos puede variar debido al hecho de que el código genético es degenerado, existiendo más de un codón para cada aminoácido, con su obvia excepción del codón de metionina. Inclusive puede compararse la evolución de genes en poblaciones separadas geográficamente pero pertenecientes a una sola especie (Choudhuri, 2014).

Por último, si las relaciones filogenéticas que se están analizando son de grupos de organismos más ampliamente divergentes, por ejemplo entre bacterias y eucariotas, lo adecuado es usar secuencias de nucleótidos con lenta evolución por ejemplo RNA ribosomal o secuencias de proteína (Lemey *et al.,* 2009; Fitch, *et al.,* 1967; Choudhuri, 2014).

En resumen la decisión de utilizar las secuencias de nucleótidos o proteínas depende de las propiedades de las secuencias y los propósitos del estudio. (Lemey, *et al.*, 2009; Fitch, *et al.,* 1967).

1.5.2.2 Alineamiento múltiple de secuencias

Según los expertos el alineamiento múltiple de secuencias es el paso más importante en la construcción de un árbol filogenético confiable, dado que se dio una más detallada explicación de los mismos en secciones previas, sólo se destacarán algunos detalles del mismo. Mediante un alineamiento múltiple de

secuencias se pueden identificar bloques conservados de residuos. Un buen alineamiento debería tener pocos espacios (gaps). Dichos gaps indican inserciones o deleciones de residuos durante la evolución. Un aspecto importante es que el usuario decide si usar el alineamiento entero o solo parte de éste, en regiones donde el alineamiento sea ambiguo, adicionalmente, una edición en tales alineamientos puede ser hecha usando Gblocks. Gblocks es un programa computacional que elimina las posiciones pobremente alineadas, así como las regiones divergentes de un alineamiento de secuencias de DNA o proteínas. Dichas posiciones pueden no ser homologas o pueden haber sido saturadas de sustituciones, a lo largo de su historia evolutiva, por lo que es conveniente eliminarlas previo a nuestro análisis filogenético. Gblocks selecciona bloques de manera similar a como era hecho de manera manual por un investigador capacitado, pero siguiendo un conjunto de condiciones reproducibles. Se puede acceder a Gblocks en http://www.phylogeny.fr/version2_cgi/one_task.cgi?task_type5gblocks o en http://molevol.cmima.csic.es/castresana/Gblocks_server.html. En el primer enlace se provee un ejemplo de cómo ingresar los datos del alineamiento; mientras que en el segundo enlace se presenta un ejemplo de un archivo de salida, mostrando los bloques seleccionados de un alineamiento de proteínas (Choudhuri 2014; Dereeper *et al.,* 2008; Hall, 2011).

1.5.2.3 Selección del modelo de evolución

Un modelo evolutivo, es un modelo de sustitución de nucleótidos o aminoácidos y por consecuencia de divergencia de las distintas secuencias. Dicho modelo, tiene la función de describir el modo y la probabilidad de que una secuencia de nucleótidos o proteínas cambie a lo largo del tiempo. Es decir, estos modelos describen para cada uno de los sitios en nuestra matriz de sustitución, la probabilidad de que se produzca el cambio de un nucleótido o aminoácido a otro a lo largo de las ramas de un árbol filogenético dado.

Estos modelos tratan de explicar la complejidad existente en aquellos procesos biológicos como la mutación, mediante el empleo de patrones simples que pueden

ser descritos y predichos usando un pequeño número de parámetros. Los modelos tratan pues de predecir la tasa de sustitución de nucleótidos o aminoácidos en un sitio dado, además de, la distribución de sustituciones a lo largo de la secuencia entera. La tasa diferencial de sustituciones a lo largo de la secuencia es conocida como la tasa de heterogeneidad, la cual se debe a que no todas las regiones de una molécula de DNA o proteínas presentan la misma tasa de cambio (Mount, 2004; Felsenstein, 2004).

El paso siguiente a un alineamiento múltiple es la elección del adecuado modelo evolutivo. Existen muchos de estos modelos estadísticos, teniendo en común que todos deben cumplir ciertas suposiciones como son a) cada posición en una secuencia de nucleótidos o aminoácidos evoluciona independientemente. Aunque este supuesto es difícil de cumplir, ya que existen los denominados hot spots de mutaciones, además, de que algunas mutaciones son más toleradas que otras, y b) la tasa de sustitución es constante en el tiempo y entre los distintos linajes (Felsenstein, 2004; Mount, 2004; Lemey *et al.,* 2009; Hall, 2011).

La manera más sencilla para determinar si dos o más secuencias han divergido es contar el número de sustituciones entre ellas. Sin embargo existen varias deficiencias en un método tan sencillo como éste. Algunos de los modelos de sustitución de aminoácidos más conocidos son el modelo de Dayhoff (PAM), el modelo Bishop Friday, el modelo Jones Taylor Thornton (JTT), el modelo Whelan and Goldman (WAG) y el modelo de Le Gascuel (LG). El modelo más simple es el de Bishop Friday, el cual asume que todos los aminoácidos se presentan con igual frecuencia y todas las sustituciones ocurren a la misma tasa. El resto de los modelos asumen diferentes frecuencias de aminoácidos y diferentes tasas de sustitución, las cuales son experimentalmente determinadas. Algunos softwares como MEGA pueden realizar los análisis pertinentes para determinar el mejor modelo de sustitución que mejor se ajuste a cada juego de datos (Le S. Q. & Gascuel, 2008; Yang *et al.,* 1998).

1.5.2.4 Construcción del árbol filogenético

Otro paso crucial en la construcción de un árbol filogenético es la elección de un método apropiado para la construcción del mismo. A la fecha existen una variedad de métodos para la construcción de árboles filogenéticos; al igual que el paso correspondiente a la elección del mejor marcador molecular, cada uno de los métodos de construcción de árboles filogenéticos, tiene sus pros y sus contras. Esta es un área muy especializada de la computación y la estadística, que escapa del objetivo del presente trabajo, por lo que solamente algunos principios generales serán descritos a continuación. Los métodos para construir árboles filogenéticos pueden ser clasificados en dos grandes grupos: métodos basados en distancia y los métodos basados en caracteres también conocidos como métodos discretos (Mount, 2004; Felsenstein, 2004).

Métodos basados en matriz de distancia.

La distancia es una medida del grado de divergencia entre dos secuencias. Utilizamos la distancia como una aproximación al tiempo desde que las dos secuencias se separaron. En este caso lo primero que se hace es generar una matriz de distancias por parejas de secuencias a partir del alineamiento múltiple. En estos métodos toda la información de similitudes y diferencias entre dos secuencias queda resumida en un único dato numérico (Mount, 2004; Felsenstein, 2004).

Existen diferentes análisis estadísticos para calcular la distancia entre dos secuencias. En teoría podríamos utilizar un análisis tan sencillo como; contabilizar las posiciones cambiadas entre dos secuencias, pero la mayoría de los análisis toman en cuenta que las posiciones pueden estar saturadas, es decir en el alineamiento vemos los cambios definitivos pero no aquellos que han afectado a la misma posición varias veces. El objetivo de un buen análisis es tener una media de distancia que se comporte linealmente con el tiempo. Debido al problema conocido como saturación llegará un momento que aunque aumente el tiempo desde que dos secuencias se separaron las mutaciones quedarán enmascaradas

por actuar sobre la misma posición y será difícil contabilizarlas para la distancia. Para solucionar este problema se suele utilizar un modelo de mutación mediante el que intentamos evaluar la distancia real en base a las mutaciones que podemos observar (Mount, 2004; Felsenstein, 2004).

Existen numerosos métodos basados en distancia, pero los más utilizados son el (UPGMA) el método de agrupación de pares no ponderado con media aritmética (the unweighted pair group method with arithmetic mean) y el Neighbor-Joining. Estos métodos basados en matrices de distancias son muy rápidos y permiten utilizar secuencias de longitud larga, así como, un elevado número de ellas con recursos computacionales de memoria y tiempo modestos (Mount, 2004; Felsenstein, 2004).

El método UPGMA emplea la matriz de distancia más simple, la cual se basa en el agrupamiento secuencial para construir un árbol filogenético enraizado. Inicialmente todas las secuencias son comparadas a través de un alineamiento de pares de secuencias para calcular la matriz de distancia. Usando esta matriz, las dos secuencias con la distancia mínima son identificadas y agrupadas como un sólo par. Luego la distancia entre este par y todas las otras secuencias es nuevamente calculada para formar una nueva matriz. Usando esta nueva matriz, la secuencia que es más cercana al primer par es identificada y agrupada. Este proceso se repite hasta que todas las secuencias han sido incorporadas en el grupo. Debido a que es un proceso no ponderado, se asume que todos los alineamientos de pares de secuencias contribuyen de igual forma (Wheeler *et al*., 2007). Resulta conveniente usar este método solo sí se cumple la hipótesis del reloj molecular, donde todas las secuencias han evolucionado a la misma velocidad, dado que genera árboles ultra métricos, es decir cuando la distancia entre todos los taxa y el ancestro común son iguales (Wheeler *et al*., 2007).

El método de Neighbor Joining es el método de matriz de distancia más usado. Este comienza con un árbol estrella, es decir, se asume que todas las ramas radian de un nodo interno formando un patrón similar al de una estrella. Luego un par de secuencias son elegidas aleatoriamente y removidas del árbol inicial, para

luego ser unidas a un segundo nodo interno el cual es conectado por una rama al centro del árbol inicial. Posteriormente la longitud de las ramas es calculada y estas dos secuencias son regresadas a su posición original en el árbol inicial, luego otro par de secuencias es tomado para repetir la misma operación. La meta es repetir todo este proceso hasta que todos los posibles pares han sido examinados y así descubrir la combinación de vecinos que minimiza la longitud total del árbol filogenético (Saitou y Nei, 1986; Gascuel *et al*., 2006).

Con este método se obtiene un árbol no enraizado y aditivo, en el cual la longitud de sus ramas es un indicativo de cambio evolutivo. Las ramas presentan diferentes distancias al punto de origen porque no asume la existencia de un reloj molecular, y por lo tanto la tasa de cambio varía entre distintos linajes y secuencias a diferencia del método de UPGMA. El método de Neighbor-Joining representa mejor la situación de la mayoría de los casos, por lo que en la actualidad se utiliza más.

Métodos basados en carácter.

En contraste a los métodos de matriz de distancia, los métodos basados en carácter utilizan la secuencia en sí mismo más que la distancia obtenida de apareamiento de pares de secuencias. Un carácter es un sitio (posición) en el alineamiento. Hay dos métodos basados en carácter muy populares, el método de máxima parsimonia (MP) y el método de máxima verosimilitud (ML) (Kolaczkowski *et al*., 2004).

El método de máxima parsimonia calcula muchos árboles del juego de datos obtenido y asigna un costo a cada árbol. La suposición de máxima parsimonia es que el árbol más simple es el árbol más plausible. Por lo tanto el árbol más simple es aquel que requiere el menor número de cambios para explicar los datos en el alineamiento (Choudhuri, 2014).

Para elegir el mejor árbol, aquel árbol que implicase menos cambios, en teoría habría que evaluar todos los árboles posibles. Dicha evaluación es viable sí el número de secuencias no es muy grande, pero a medida que este número

aumenta el número de árboles crece desmesuradamente, haciendo poco factible evaluar todos los árboles obtenidos. En la práctica los programas utilizan un método heurístico para evaluar sólo los árboles más razonables, desechando directamente los más improbables (Farris, 1970; Fitch, 1971).

El resultado final de un algoritmo de máxima parsimonia no tiene por qué ser un único árbol puesto que puede que existan varios árboles que impliquen un mismo número de mutaciones (Farris, 1970; Fitch, 1971).

El método de máxima verosimilitud busca el árbol más probable que haya sido generado a partir de los datos de los que disponemos. En este método partimos de los datos y de un modelo de evolución. Partiendo de esta base se calcula la probabilidad de que nuestros datos hayan sido generados por los distintos árboles posibles y se devuelve el árbol que presenta una máxima probabilidad (Yang, 1996; Whelan *et al*., 2001).

Para poder calcular uno de estos árboles es imprescindible elegir un modelo de mutación *a priori*. Una vez determinado el modelo, el algoritmo hará una estimación maximoverosímil de los parámetros relativos a las tasas de mutación así como del árbol (Yang, 1996; Whelan *et al,*. 2001).

De manera análoga al método de máxima parsimonia, en éste caso para calcular el árbol más verosímil también deberían explorarse todos los árboles posibles, pero dado que eso es computacionalmente imposible de llevar a cabo también se utilizan métodos heurísticos para explorar tan sólo los árboles más razonables (Yang, 1996; Whelan *et al*., 2001).

Este método, según los expertos, tiene la ventaja frente al de máxima de parsimonia de utilizar con mayor eficiencia la información filogenética contenida en el alineamiento múltiple. Es decir, dado un mismo alineamiento éste método tiende a generar un resultado más cercano a la realidad (Yang, 1996; Whelan *et al*., 2001).

La desventaja principal del método es el coste computacional que conlleva. De hecho, dos aspectos importantes a la hora de usar este método son; el número de secuencias y la longitud de las mismas. A menos que se disponga de redes computacionales, el empleo del método de máxima verosimilitud en computadoras comerciales está muy restringido a estos dos parámetros (Yang, 1996; Whelan *et al*., 2001).

1.5.2.5 Evaluación de la confiabilidad de un árbol filogenético

Dado lo leído en párrafos anteriores sobre los pros y contra de elegir un marcador molecular y un método específico de construcción de árboles filogenéticos era de esperarse que se evaluara la fiabilidad del mismo. Determinar la fiabilidad de un árbol significa determinar si la topología del árbol es correcta o si un mejor árbol puede ser obtenido. Estas incógnitas son resueltas mediante el método de bootstrap de reconstrucción de árboles. El método de bootstrap es una técnica estadística de remuestreo con reemplazamiento, que funciona de la siguiente manera: a partir de nuestro alineamiento original se generan subjuegos de pseudomuestras tomando aleatoriamente una porción de secuencias del alineamiento original. En este muestreo con reemplazamiento se escogen secuencias al azar para generar las pseudomuestras, pudiendo incluir una misma secuencia varias veces mientras que otras pueden nunca ser incluidas. Una vez que se han generado un cierto número de pseudomuestras, por ejemplo entre 500 a 1000 muestras del juego original de secuencias, se infiere un árbol filogenético con el mismo análisis que se empleó en la muestra original, obteniendo así muchos árboles de bootstrap. El conjunto de árboles obtenidos a partir de las pseudomuestras se resume en un árbol final que conocemos como árbol de consenso. Cuando el patrón de ramificación de una rama interior (topología de la rama) en el árbol original es reproducida en el árbol de bootstrap, a esta rama se le asigna un valor de 1 (valor de identidad). En otras palabras, cuando a una rama se le asigna un valor de 1 se asume con precisión que se trata de un clado con dos taxa hermanos, como se refleja en el árbol original y en el de bootstrap.

Contrariamente, cuando el patrón de ramificación de una rama interna en el árbol original no es reproducido en el árbol de bootstrap, a dicha rama se le asigna un valor de 0. Este proceso se repite cientos de veces, según el número de pseudomuestras tomadas y el porcentaje de veces en que a cada rama interior se le da un valor de 1 se calcula. Este es conocido como valor de bootstrap o valor de confianza de bootstrap. Como regla general, si el valor de confianza de bootstrap para una rama interior es de 95% o superior, la topología de la rama es considerada correcta. Los valores de bootstrap, expresados como porcentajes, son indicados sobre las ramas. Por consiguiente, un valor de bootstrap de 95 indicado sobre la rama significa que el 95% de los árboles de bootstrap apoyan la topología de la rama obtenida en el árbol filogenético consenso (Felsenstein, 1985; Hillis *et al.,* 1993; Lake *et al.,* 1998).

La base estadística del análisis de bootstrap es que si un parámetro puede estimarse a partir de muestras extraídas de una población, entonces la fiabilidad de la estimación de este parámetro puede ser verificada mediante la extracción de nuevas muestras de la misma población. Cuanto mayor sea el número de muestras, mayor es el nivel de confianza de la estimación (Felsenstein, 1985; Hillis *et al.,* 1993; Lake *et al.,* 1998).

2 Proteínas homólogas evolucionaron de un ancestro común

La evolución es un concepto unificador en bioquímica. La estructura de las proteínas que existen hoy ha sido moldeada por la evolución. Se puede decir que la evolución proteica es la exploración del espacio por parte de secuencias de aminoácidos en busca de alguna variante selectivamente ventajosa. La evolución actúa a nivel de función de proteínas, en un ciclo de retroalimentación que selecciona las secuencias génicas. Las relaciones evolutivas observadas entre estructuras proteicas iluminan las relaciones existentes entre secuencia-estructura-función (Akashi, 2012). Proteínas que están relacionadas por evolución tienen estructuras similares, esto último se debe a que las mutaciones actúan

sobre la secuencia, mientras que la evolución actúa sobre la estructura y función. Se puede decir que las proteínas evolucionan a varias escalas:

- Dominios individuales son sujetos a sustitución de aminoácidos, inserciones y deleciones. La mayoría, aunque no todas de tales mutaciones producen cambios mínimos en la estructura y modificaciones menores de la función. Sin embargo, el efecto acumulativo de varias de estas mutaciones locales puede producir grandes cambios en la estructura de los distintos dominios, y en el desarrollo de una nueva función (Rigden, 2009).
- En un nivel estructural superior, las proteínas pueden evolucionar por "mezclado y combinado" (mixing and matching) de dominios enteros. Dado que la mayoría de las proteínas contienen más de un dominio y/o más de una cadena polipeptídica. La evolución ha recombinado un repertorio relativamente pequeño de dominios en muchas diferentes asociaciones. Algunas veces los dominios individuales contribuyen siendo componentes funcionales especializados de la estructura general. Por ejemplo, el dominio de unión a $NADH^{+}$ el cual está presente en muchas deshidrogenasas que comparten dicho cofactor, en combinación con dominios catalíticos de muy distintas estructuras y funciones, específicos para diferentes sustratos. En otros casos, la función de la proteína como un todo varía más ampliamente con la combinación de dominios (Letunic *et al.,* 2012).

Los retos de estudiar la evolución proteica incluyen varias metas relacionadas, como son:

- Un método confiable para definir cuales proteínas son homólogas.
- Una comprensión de las vías evolutivas admisibles que conectan a las proteínas homólogas.
- Una comprensión de que limita la divergencia de secuencias proteicas, estructuras y funciones.

Por definición dos o más proteínas son homólogas si ellas son descendientes de un ancestro común. Debido a que muy difícilmente podemos observar homología

directamente, nosotros debemos inferir homología de similitud en secuencia, estructura o función. Teniendo esto dos implicaciones (Rigden, 2009).

Se debe calibrar nuestras medidas de similitud para proveer umbrales para una conclusión confiable acerca de la homología. Para excluir no-homólogos, que no obstante, muestran algún grado de similitud, nosotros debemos establecer criterios lo suficientemente estrechos para poder discriminar entre ambas posibilidades (Rigden, 2009; Letunic *et al.,* 2012).

Sería muy útil saber de cuán lejos proviene la divergencia evolutiva. Esto nos permitiría evaluar si la elección de un criterio estricto para inferencia de similitud de homología nos lleva a perder muchos homólogos genuinos que son altamente divergentes. Un método para demostrar que dos proteínas muy distintas A y B son homólogas, sería exhibir una serie de proteínas intermedias que formasen una vía conectada entre A y B, tal que cada sucesivo par de la ruta presenta alta similitud, por desgracia esto no siempre es posible (Rigden, 2009; Reid, 2007).

En la evolución de dominios proteicos individuales, una sucesión de mutaciones lleva a divergencia progresiva de secuencias y estructuras. Los criterios para reconocer homólogos son extrapolaciones de comparaciones de proteínas cercanamente relacionadas en estructura y función donde el aspecto de homología no está en duda. Por ejemplo, el cachalote y la foca presentan una mioglobina que tiene 127/153= 83% de residuos idénticos mostrado así en un alineamiento de secuencias, tienen una estructura muy similar, incluyendo el mismo cofactor unido por los residuos correspondientes, además de cumplir la misma función. Por lo que el carácter de homología entre ambas secuencias no se cuestiona (Rigden, 2009; Reid, 2007; Letunic *et al.,* 2012).

Continuando con este mismo ejemplo de las globinas y extrapolándolo a otros grupos de mamíferos se ha encontrado que la estructura diverge más lentamente que la secuencia. Por lo que, para parientes distantes la homología es más fácil de reconocer en la estructura más que en la secuencia. En la mayoría de los casos el patrón de plegamiento básico de una familia permanece relativamente igual.

Usualmente hay al menos un pequeño juego de residuos conservados al menos en carácter fisicoquímico, a pesar de que esto requiere conocimiento de la estructura para conocer el papel que juega sobre la arquitectura del plegamiento (Rigden, 2009; Reid, 2007; Letunic *et al*., 2012). Esto junto con características funcionales comunes, da mayor certeza a la hora de discernir entre secuencias homólogas y no homólogas.

Lo que sigue creando dificultades son:

- Casos en los cuales los plegamientos son tentativamente similares en ausencia de cualquier similitud de secuencia arriba del nivel umbral fijado y que a su vez, no presentan función análoga. El caso de la ficocianina y las globinas son un ejemplo de esto.
- Casos de secuencias similares con diferentes patrones de plegamiento.

Afortunadamente este tipo de casos son muy raros.

2.1 Clasificación de proteínas

Las proteínas son el principal componente de los organismos y como tal están sujetas a selección natural, la cual promueve nuevas características adaptativas o remueve mutaciones deletéreas. Estas adaptaciones, las cuales pueden ser en respuesta a un ambiente interno o externo, oscilan de sustituciones en un solo aminoácido a inserciones y deleciones bastante grandes de residuos, pudiendo resultar algunas veces en la adición de meros ornamentos estructurales a los núcleos evolutivamente conservados de dominios estructurales, de ahí la finalidad de clasificarlas. Siendo la superfamilia una de las categorías más usadas para las distintas proteínas (Rigden, 2009; Reid, 2007).

Una superfamilia es un conjunto de proteínas que se piensa están relacionadas evolutivamente. Consecuentemente, pueden ser inferidas relaciones de superfamilia mediante similitud de secuencia, la cual debería ser detectada utilizando cualquiera de los métodos de alineamiento de secuencias tradicionales o mediante búsquedas más sensibles que incluyen modelos ocultos de Markov.

En ausencia de similitud de secuencia, también puede ser descubierta homología remota a partir de la similitud de estructura y/o función. Aunque, por esta vía no existe un enfoque ampliamente aceptado para evaluar si una semejanza estructural o funcional es estadísticamente significativa. Debido a eso, los límites utilizados para definir las relaciones de superfamilia pueden ser arbitrarios e inclusive algo subjetivos. Hoy en día, las distintas bases de datos han usado definiciones estándar y ampliamente aceptadas de lo que es una superfamilia. Sin embargo, en todas ellas, cierto grado de subjetividad permanece a la hora de asignar una proteína dentro de una superfamilia (Rigden, 2009; Reid, 2007; Csaba *et al*., 2009).

El concepto de familia es aún más difícil de definir. Hoy en día, una familia se entiende comúnmente como un conjunto de proteínas homólogas que cumplen con ciertos criterios. Por ejemplo, una familia de secuencias en un nivel particular de similitud de secuencia, agrupa juntas todas aquellas proteínas que comparten al menos este nivel de similitud de secuencia; una familia funcional agrupa juntas homólogos que tienen la misma función; una familia ortóloga agrupa ortólogos; etc. Por lo tanto, la definición de familia variará dependiendo del enfoque de cada una de las bases de datos existentes (Rigden, 2009; Csaba *et al*., 2009).

2.1.1 Superfamilia de proteínas

Una superfamilia proteica es la agrupación más grande de proteínas para la cual una relación de parentesco puede ser inferida. Usualmente esta relación de homología o ancestría común está basada sobre alineamientos estructurales, recordando que la estructura esta generalmente mejor conservada que la secuencia en la evolución, por lo que muchas proteínas despliegan características estructurales comunes, además de similitud mecanística, a pesar inclusive de que la similitud de secuencia no es evidente (Todd *et al*., 2001; Jiang *et al.,* 2007). Las superfamilias normalmente se componen de varias familias proteicas, las cuales muestran similitud de secuencia dentro de su respectiva familia, pudiendo carecer

de similitud de secuencia entre miembros de las distintas familias (Todd *et al*., 2001).

Las superfamilias de proteínas son identificadas usando una variedad de métodos. Por ejemplo, miembros cercanamente relacionados pueden ser identificados por diferentes métodos a aquellos necesarios para agrupar los miembros más divergentes evolutivamente. Siendo los tres principales métodos: la similitud de secuencia, la similitud de estructura y la similitud mecanística o funcional (Todd *et al*., 2001; Rigden, 2009).

Similitud de secuencia: este es el método históricamente más usado para inferir homología. Dicho método es considerado un buen predictor de parentesco, ya que es más probable que un grupo de secuencias parecidas, resulten de duplicación génica y su posterior divergencia, que del resultado de convergencia evolutiva. Las secuencias de aminoácidos están más conservadas que las secuencias de DNA, debido a que el código genético es degenerado, por lo que estas últimas son más sensibles a este método de detección. Dentro de la misma secuencia proteica encontramos unas regiones más conservadas que otras, por ejemplo, regiones funcionalmente importantes como sitios catalíticos y sitios de unión son regiones consideradas menos tolerantes a cambios en sus secuencias. El uso de similitud de secuencia para inferir relaciones de ancestría común dentro de una superfamilia tiene varias limitaciones. Ya que no existe un nivel mínimo de similitud de secuencia garantizado para producir una estructura idéntica. Por otro lado, durante largos periodos de tiempo evolutivo, las proteínas relacionadas pueden casi carecer de similitud de secuencia detectable entre sí. Otro caso sería que secuencias con muchas inserciones o deleciones resultan difíciles de alinear y por lo tanto identificar regiones homólogas entre ellas. No obstante, todas estas desventajas, dicho método es el más comúnmente usado para evidenciar o inferir parentesco, ya que el número de secuencias conocidas supera con creces el número de estructuras terciarias conocidas. En ausencia de información estructural, la similitud de secuencia es el principal criterio por el cual las proteínas

pueden ser asignadas a una superfamilia específica (Rison *et al*., 2002; Rigden, 2009).

Similitud estructural: este método se basa en que la estructura está mejor conservada evolutivamente que la secuencia, de tal forma que proteínas con estructuras muy similares pueden tener secuencias completamente diferentes. Ya que en periodos de tiempo evolutivo muy largos, muy pocos aminoácidos muestran conservación en su secuencia, sin embargo, elementos de estructura secundaria y motivos estructurales terciarios se encuentran altamente conservados. Por consiguiente, la estructura terciaria de una proteína puede ser usada para detectar homología entre proteínas aun cuando no exista evidencia de parentesco entre sus secuencias (Rison *et al*., 2002; Rigden, 2009).

Similitud mecanística o funcional: este método se basa en el hecho de que el mecanismo catalítico de las enzimas dentro una superfamilia está muy conservado, aunque la especifidad de sustrato puede ser significativamente diferente. Además, aquellos residuos involucrados en la reacción tienden a encontrarse en el mismo orden en la secuencia proteica. Sin embargo, el mecanismo por sí solo no es suficiente para inferir relación de parentesco, ya que algunos mecanismos han evolucionado de manera convergente múltiples veces de manera independiente, y por lo tanto forman distintas superfamilias (Rison *et al*., 2002; Rigden, 2009).

Las superfamilias proteicas representan el límite actual de nuestra habilidad para identificar ancestría común. Además, de ser la agrupación evolutiva basado en evidencia directa que puede ser reconocido, por lo tanto, el origen de las superfamilias está entre los eventos evolutivos más antiguos actualmente estudiados. Algunas superfamilias tienen miembros presentes en todos los reinos de la vida, indicando que el último ancestro común de tal superfamilia estuvo en el último ancestro común universal de toda la vida (LUCA) (Rison *et al*., 2002; Rigden, 2009).

Dado que la mayoría de las proteínas contienen múltiples dominios, en el caso de eucariontes entre el 66 y 80% de sus proteínas están en esta situación, mientras que del 40 al 60% de las proteínas procariontes presentan múltiples dominios. Con el paso del tiempo muchos de los dominios de las distintas superfamilias se han mezclado entre sí. De hecho, es muy raro encontrar superfamilias totalmente aisladas, por último ya que el número de combinaciones de dominios visto en la naturaleza es pequeño comparado con el número de posibilidades, esto sugiere que la selección actúa sobre todas las combinaciones (Rigden, 2009).

2.1.2 Familia de proteínas

Una familia de proteínas es un grupo de proteínas relacionadas evolutivamente. Por lo tanto las proteínas dentro de una familia descienden de un ancestro común, teniendo normalmente una estructura tridimensional parecida entre los miembros de una familia determinada, además, de compartir una función y de presentar similitud de secuencia de manera significativa. La más importante de estas características es precisamente la similitud de secuencia, ya que este aspecto es el indicador más estricto de homología y por ende el más claro indicador de ancestría común. Por lo que actualmente existen protocolos muy bien desarrollados para evaluar sí la similitud entre un grupo de secuencias usando métodos de alineamiento es significativa. Esto se basa en el principio de que un grupo de proteínas que no comparten un ancestro común, es muy difícil que muestren similitud de secuencias de manera significativa, teniendo en los alineamientos de secuencias una poderosa herramienta para identificar a los distintos miembros de una familia de proteínas (Kunin *et al*., 2003; Dayhoff, 1975; Chothia *et al*., 1986).

Como ya se mencionó antes en este mismo trabajo, las familias de proteínas son agrupadas juntas en categorías más grandes llamadas superfamilias de proteínas basados casi exclusivamente en criterios de similitud estructural y similitud mecanística o de función, esto incluso si no es posible identificar similitud de secuencia (Kunin *et al*., 2003).

Actualmente, más de 60 000 familias proteicas han sido definidas, aunque la ambigüedad en la definición de familia proteica, tema que se ha abordado someramente líneas arriba en este trabajo, hace que muchos expertos piensen en un número bastante mayor al aquí citado (Kunin *et al*., 2003; Rigden, 2009).

Como ya se había mencionado el uso de familia proteica depende del contexto, ésta puede indicar grandes grupos de proteínas con el nivel más bajo posible de similitud de secuencia o por otro lado puede indicar un grupo muy estrecho de proteínas con casi idénticas secuencias, función y estructura tridimensional, o bien cualquier grupo entre estos dos extremos. Para distinguir entre estas situaciones, el termino superfamilia de proteínas se usa seguido para proteínas distantemente relacionadas, de quienes el parentesco no es detectable por similitud de secuencia, pero si mediante las características estructurales compartidas. Hay quienes concuerdan en que las superfamilias, grupos de proteínas que presentan homología estructural, contienen a las familias, grupo de proteínas que presentan homología de secuencia, las cuales a su vez contienen otro grupo denominado subfamilias proteicas (Chothia *et al*., 1986; Kunin *et al*., 2003; Gandhimathi *et al,*. 2011).

El concepto de familia proteica fue creado hace mucho tiempo, cuando muy pocas estructuras proteicas e inclusive secuencias eran conocidas, por estas fechas, debido a limitaciones en las distintas técnicas de purificación y cristalización de estructuras proteicas, solamente pequeñas proteínas de un solo dominio eran estructuralmente conocidas. Desde entonces, y con el perfeccionamiento de las distintas técnicas, se encontró que muchas proteínas se conforman de múltiples unidades que son estructural y funcionalmente independientes conocidas como dominios. Debido al barajeado de dominios, diferentes dominios en una proteína han evolucionado independientemente. Esto ha conducido, en años recientes, a enfocarse sobre familias de dominios proteicos. Existiendo actualmente una variedad de recursos en línea dedicados a identificar y catalogar tales dominios. (Bateman *et al*., 2010; Björklund *et al*., 2005; Björklund *et al*., 2006).

Dado que las distintas regiones de cada proteína tienen diferentes limitaciones funcionales (características críticas a la estructura y función de la proteína). Un ejemplo claro es el sitio activo de una enzima el cual requiere que ciertos aminoácidos estén orientados de una manera tridimensional especifica. Por otro lado, la intercara de unión proteína-proteína puede consistir de una gran superficie cuyas limitantes están sobre la hidrofobicidad o polaridad de aminoácidos específicos. Ya que las regiones funcionalmente limitadas de las distintas proteínas evolucionan más lentamente que aquellas regiones que no tienen tales limitaciones, se han originado bloques de secuencias conservadas fácilmente discernibles cuando las secuencias de una familia proteica son comparadas en un alineamiento de secuencias. Tales bloques reciben el nombre de motivos (Björklund *et al*., 2005; Björklund *et al*., 2006).

2.2 Clasificación de enzimas

Una de las maneras de clasificar a las distintas proteínas es en base a su función; debido a esto existen bases de datos y sistemas de anotaciones que están disponibles para la descripción de la función molecular de proteínas, las cuales son muy útiles para el estudio de relaciones estructura función. Quizás uno de los más antiguos sistemas para describir la función molecular de proteínas es el esquema de números de la Comisión de Enzimas (EC). Los números EC del inglés (Enzyme Commission numbers) son un esquema de clasificación numérica para las enzimas, basado en las reacciones químicas que catalizan (Webb 1992; Rigden, 2009).

Como sistema de nomenclatura de enzimas, cada número EC está asociado a un nombre recomendado para dicha enzima. Estrictamente hablando los números EC no especifican enzimas, sino que en realidad, los números EC codifican reacciones catalizadas por enzimas. Enzimas diferentes (por ejemplo que procedan de organismos diferentes) que catalicen la misma reacción recibirán el mismo número EC. Además, a través de la evolución convergente, proteínas completamente diferentes pueden catalizar una reacción idéntica y por lo tanto

tendrían asignado el mismo número EC, estas son denominadas enzimas isofuncionales no homólogas (Omelchenko *et al*., 2010).

Cada código de enzimas consiste en las dos letras EC seguidas por 4 números separados por puntos. Estos números representan una clasificación progresivamente más específica, cuadro 8 (Moss 2006; Omelchenko *et al*., 2010).

Por ejemplo, la enzima tripéptido aminopeptidasa tiene el código EC 3.4.11.4, construido así:

EC 3 enzimas hidrolasas (enzimas que usan el agua para romper alguna otra molécula).

EC 3.4 son hidrolasas, que actúa sobre los enlaces peptídicos.

EC 3.4.11 son aquellas hidrolasas que cortan el aminoácido en el extremo amino terminal de un polipéptido.

EC 3.4.11.4 son aquellas que actúan sobre el amino terminal de un tripéptido.

Cuadro 8. Clasificación de acuerdo al primer número EC.

Grupo	Reacción catalizada	Reacción típica	Ejemplo de enzima
EC 1 Oxidorreductasas	Reacciones de oxidación/reducción y de transferencia de átomos de H, O o electrones desde una substancia a otra.	AH + B → A + BH (reducido) A + O → AO (oxidado)	Deshidrogenasa, oxidasa
EC 2 Transferasas	Transferencia de un grupo funcional desde una substancia a otra. El grupo puede ser metil-, acil-, amino- o fosfato.	AB + C → A + BC	Transaminasa, quinasa
EC 3 Hidrolasas	Formación dos productos de un substrato por hidrólisis.	AB + H2O → AOH + BH	Lipasa, amilasa, peptidasa
EC 4 Liasas	Adición o eliminación no hidrolítica de grupos de los substratos. Pueden romper los enlaces C-C, C-N, C-O o C-S.	RCOCOOH → RCOH + CO2	Descarboxilasa
EC 5 Isomerasas	Isomerización de una molécula.	AB → BA	Isomerasa, mutasa

EC 6 Ligasas	Unión de dos moléculas por síntesis de nuevos enlaces C-O, C-S, C-N o C-C con la rotura simultánea de ATP.	X + Y+ ATP → XY + ADP + Pi	Sintetasa

Con el fin de superar las limitaciones atribuidas a la clasificación EC, se han creado nuevas bases de datos destacando dos que recientemente se han establecido para clasificar las enzimas y sus reacciones: la primera es EzCatDB (Enzyme catalytic-mechanism Database) y la segunda MACiE (Mechanism, Annotation and Classification in Enzymes). Ambas bases de datos se centran en la descripción y clasificación de los mecanismos de reacción enzimática en lugar de las propias reacciones, ya que se ha argumentado que una clasificación basada en la reacción como el sistema de la EC no refleja relaciones de homología, pudiendo adquirir dos enzimas completamente distintas la misma función, por ejemplo la función de alcohol deshidrogenasa en animales observada en dos superfamilias distintas, pero agrupadas por el sistema EC en el mismo grupo, EC 1.1.1.1 (Nagano *et al*., 2007; Holliday *et al*., 2007). Complementario a estas bases de datos, existen Atlas de sitios catalíticos que proveen información detallada de los residuos aminoácidos específicos que directamente participan en los mecanismos catalíticos para enzimas de estructura conocida. Varias bases de datos proveen mayor descripción de todos los residuos de aminoácidos de una proteína involucrados en la unión de moléculas biológicamente importantes tal como sustratos o cofactores. Otros sistemas de anotaciones ampliamente usados para la función de proteínas incluyen a KEGG, el cual inicialmente tuvo como objetivo la descripción de las rutas metabólicas así como redes de reacciones biológicas; la cual se ha extendido ahora a un sistema de funciones biológicas. Por otro lado tenemos a FUNCAT (the Functional Catalogue), que clasifica las funciones de las proteínas en un árbol jerárquico único. Las bases de datos KEGG y FUNCAT han estado tradicionalmente más enfocados sobre los procesos biológicos en los cuales las proteínas están involucradas que en sus actividades moleculares, pero ambas bases de datos, sin embargo, pueden proveer pistas muy útiles con respecto a lo que se conoce como función molecular (Nagano *et al*., 2007; Holliday *et al*., 2007; Ruepp *et al*., 2004).

2.3 Divergencia de la función durante la evolución proteica

Estudios recientes han presentado muchos nuevos e interesantes puntos de vista sobre la evolución de proteínas y han descrito varias herramientas para estimar y comprender las fuerzas evolutivas que actúan sobre las proteínas. Cuando la divergencia funcional ha ocurrido, la selección positiva puede ser detectada a nivel de nucleótidos. Estas adaptaciones pueden ocurrir en una única posición o en grupos de sitios de manera coevolutiva. Por ejemplo se ha observado que los aminoácidos en el núcleo de una proteína tienden a evolucionar si la superficie evoluciona también, como si ambas estuviesen conectadas. En un nivel superior, un gran número de variaciones son observadas entre homólogos remotos en algunas superfamilias. Mientras que la especificidad de sustrato puede ser diversa a través de una superfamilia, la reacción química se mantiene más a menudo dentro de ésta. En muchas superfamilias de enzimas, la posición de residuos catalíticos puede variar a pesar de realizar papeles funcionales equivalentes en proteínas relacionadas. Las implicaciones de la diversidad funcional dentro de las superfamilias es un foco para algunos de los proyectos de genómica estructural (Rigden, 2009; Rost, 2002).

La comprensión de los mecanismos que conducen la evolución proteica pueden ayudar en la comprensión de como los genomas son moldeados por selección natural, y más prácticamente, pueden ayudar en predecir los efectos de mutaciones sobre proteínas blanco de drogas (Rigden, 2009; Rost, 2002).

El enfoque tradicional para describir una proteína de función desconocida es buscar homología entre esta proteína y otras proteínas previamente caracterizadas, pudiendo usar las anotaciones funcionales de estas últimas sobre la primera, esto tiene como fundamento que las proteínas que descienden de un ancestro común deben compartir algún grado de funcionalidad. Aunque ahora es bien sabido que éste enfoque es propenso a errores (Rigden, 2009; Rost, 2002; Goldstein, 2008).

2.3.1 Diversidad de función a nivel de superfamilias

Las secuencias de proteínas clasificadas en la misma superfamilia han divergido algunas veces más allá de niveles que pueden ser detectados por métodos de alineamiento de secuencias estándar. A pesar de que generalmente se acepta que la estructura tridimensional de una proteína esta por mucho más conservada que la secuencia lineal durante el proceso de evolución, grandes diferencias pueden ser observadas entre estructuras de homólogos remotos. Tales diferencias estructurales pueden surgir de inserciones o deleciones (indels) de largos elementos de estructuras secundarias, inclusive de varios de estos elementos (Rigden, 2009). Un estudio reciente de indels entre estructuras homologas mostró que es normal que sucesivas inserciones de estructuras secundarias ocurran en el mismo lugar del plegamiento de una proteína durante el proceso de evolución, dando así origen a los llamados indels anidados (nested indels). Otro análisis de inserciones dentro de las superfamilias de CATH reveló que no sólo las estructuras secundarias insertadas presentan una tendencia a colocalizar en un plegamiento determinado, sino que tales inserciones seguido pueden ocurrir cerca a regiones importantes funcionalmente, tales como; sitios catalíticos de una enzima o intercaras proteina-proteina. Esta observación indica que existe una correlación entre cambios estructurales y funcionales. (Reeves *et al*., 2006; Jiang, *et al*., 2007; Letunic *et al*., 2012).

Las inserciones de nuevos elementos de estructura secundaria cerca del sitio activo muy probablemente cambiaran la función, aunque cambios más sutiles como sustitución de residuos de importancia catalítica también podrían resultar en diferencias funcionales, a pesar de que éstos son más propensos a distinguir homólogos relativamente cercanos que homólogos remotos a nivel de superfamilia (Reeves *et al*., 2006; Todd *et al*., 2001; Letunic *et al*., 2012).

A largo plazo, los procesos evolutivos a través del cual la función puede divergir entre homólogos son numerosos y difíciles de resumir. Sin embargo, en un esfuerzo reciente por comprender y categorizar tales procesos, Bashton y Chothia

han descrito e ilustrado un sub juego de éstos para comprender como la función de dominios homólogos puede cambiar dependiendo de en qué contexto se encuentren, ya sea proteína de un solo dominio o combinados con otros en proteínas de múltiples dominios (Bashton *et al*. 2007). Ejemplos de los procesos identificados incluyen casos donde la función del dominio es modificada por su combinación con otros dominios que modifican su especificidad de sustrato o casos donde la fusión de dominios resulta en proteínas multifuncionales, en las cuales cada dominio es responsable de una función particular. La aparición de cambios estructurales en las cercanías de regiones funcionales apunta hacia la diversidad funcional que es de esperarse entre miembros de una superfamilia. Y de hecho, resulta de varios estudios que indican que los homólogos remotos dentro de una superfamilia seguido realizan distintas funciones. La mayoría de estos están enfocados en la evolución de la función dentro de superfamilias particulares que generalmente muestran una diversificación funcional excepcional, un destacado ejemplo incluye a la superfamilia de las haloacido dehalogenasas y las deshidrogenasas/ reductasas de cadena corta. El estudio de estos diferentes grupos de proteínas ha revelado una gran variedad de procesos por los cuales la función diverge entre parientes, resumiendo a continuación algunos de estos procesos (Todd *et al*., 2001; Letunic *et al*., 2012).

2.3.1.1 Superfamilias funcionalmente diversas

Un subjuego de muchas familias estudiadas constituye el núcleo de los datos en la base de datos Structure Function Linkage Database (SFLD) y a pesar de su diversidad funcional, y respetando los criterios de inclusión en SFLD, todos los miembros de estas superfamilias comparten un atributo funcional común en las diversas reacciones que estas catalizan (Todd *et al*., 2001; Glasner *et al*., 2006; Baier *et al*., 2016).

El SFLD está específicamente dirigido a describir tales superfamilias de enzimas con diversidad funcional y provee una clasificación de enzimas evolutivamente relacionadas, la cual se basa sobre similitudes en sus mecanismos funcionales. Por ejemplo, el SFLD agrupa en la superfamilia de haloácido dehalogenasas,

enzimas que pueden procesar una vasta variedad de sustratos, pero que siempre actúan vía la formación de un intermediario enzima-sustrato a través de un ácido aspártico conservado, el cual a su vez facilita la escisión de enlaces C-Cl, P-C o P-O. La superfamilia haloácido dehalogenasas contiene 1, 285 secuencias clasificadas en 20 diferentes familias, cada una de las cuales cataliza una reacción única. (Todd *et al*., 2001; Glasner *et al*., 2006; Baier *et al*., 2016).

Actualmente, el SFLD solo cubre seis superfamilias. Sin embargo, la conservación de partes de la reacción química dentro de las superfamilias, parece muy común, observándose en 22 de las 31 superfamilias estudiadas por Todd y colaboradores en 2001. En contraste, la especificidad de sustrato no estuvo conservada en 20 de estas superfamilias (Todd *et al*., 2001; Glasner *et al*., 2006; Baier *et al*., 2016).

La presencia de un paso funcional común en superfamilias funcionalmente diversas sugiere que las enzimas en estas superfamilias han mantenido aspectos de su mecanismo catalítico en el curso de su diversificación evolutiva. Tal situación da un indicio de un escenario evolutivo en el cual las enzimas evolucionan nuevas funciones vía duplicación y reclutamiento, mediante el mantenimiento (aunque de forma parcial) de los mecanismos de reacción en vez de la especificidad de sustrato, dando como resultado superfamilias funcionalmente diversas observadas hoy en día (Baier *et al*., 2016).

2.3.1.2 Superfamilias con diversa especificidad

Un escenario alternativo para la evolución divergente de funciones enzimáticas dentro de las superfamilias es aquel en la cual una enzima ancestral con poca especificidad se duplica y las copias descendientes se especializan en unir sustratos más específicos. En tal escenario, la especificidad de sustrato es el factor dominante para la evolución de la función dentro de la superfamilia. En el análisis de superfamilias hecho por Todd, éste reveló que en la mayoría de los casos, los mecanismos de reacción estuvieron más conservados que la especificidad de sustrato entre enzimas homólogas. De 28 superfamilias que estuvieron involucradas en la unión de sustrato diez, no presentaron conservación por el

sustrato, inclusive otras diez tuvieron muy variados sustratos con sólo un pequeño aspecto químico en común como lo es el enlace peptídico (Todd *et al*., 2001; Glasner *et al*., 2006).

La expectativa de que la especificidad de sustrato puede estar conservada entre enzimas homólogos en una superfamilia deriva de la propuesta de Horowitz de la evolución reversa de las vías metabólicas (the backward evolution of metabolic pathways). Esta hipótesis sugiere que cuando el sustrato de una enzima se agota, un organismo que posea una nueva enzima que es capaz de producir este sustrato a partir de un compuesto precursor el cual está disponible, tendrá una ventaja selectiva sobre otros, por lo tanto la nueva enzima será fijada por la evolución dando así origen a una ruta metabólica de dos pasos. Posteriormente, un proceso evolutivo similar podría tomar lugar para los otros pasos de la ruta existente. De acuerdo con este escenario, la evolución de las rutas metabólicas va hacia atrás en comparación con la dirección del flujo metabólico. (Rison *et* al., 2002). Debido a que la enzima original tiene la capacidad de unir un sustrato el cual es el mismo que el producto de la nueva enzima, se ha sugerido que esta propiedad común puede ser usada como base para la evolución de la última enzima. Siguiendo esta idea, todas las enzimas dentro de una ruta metabólica serían homólogas, y la enzima que cataliza el paso final de la ruta sería la más antigua. Adicionalmente, la evolución de estas enzimas habría sido impulsada por la selectividad de sustrato, resultando en una tendencia de las superfamilias existentes a mostrar similitud en la especificidad de sustrato. Un ejemplo claro está en la ruta de síntesis del triptófano, en la cual varias enzimas que catalizan pasos secuenciales son claramente homólogos (Todd *et al*., 2001; Glasner *et al*., 2006). Sin embargo, el resultado de varios estudios sugiere que este hipotético proceso ha de hecho jugado un papel marginal en la evolución del metabolismo, el cual en vez habría resultado, en su mayoría, del reclutamiento de enzimas entre las distintas rutas. En realidad, las superfamilias en las cuales la selectividad de sustrato esta conservada son poco comunes, en comparación con aquellos casos donde el mecanismo catalítico esta conservado. (Todd *et al*., 2001; Rison *et al*., 2002; Rigden, 2009).

2.3.1.3 Cambios funcionales debidos a cambios en el contexto ambiental.

Los cambios funcionales entre copias duplicadas de una proteína pueden originarse no solo de cambios de la proteína en sí misma, sino más bien de cambios en las condiciones ambientales en las cuales las diferentes copias están activas. Por ejemplo, el reclutamiento de una proteína en un nuevo lugar de un organismo puede teóricamente resultar en su encuentro con pequeñas moléculas que no estaban presentes en el ambiente original de la proteína ancestral, por lo tanto, la proteína reclutada puede presentar una inesperada capacidad para unir esos nuevos ligandos disponibles. Igualmente, la función molecular de una proteína puede cambiar si otras proteínas en su ambiente sufren mutaciones, las cuales resultan en la posibilidad de nuevas interacciones, o por el contrario, que alguna interacción proteína-proteína ya no sea posible (Ojha *et al*., 2007).

2.4 ¿Cómo las proteínas desarrollan nuevas funciones?

Los mecanismos de evolución proteica que producen una nueva función o una alterada incluyen: (1) divergencia, (2) reclutamiento, y (3) mezclado y combinado de dominios.

Divergencia.

En familias de proteínas cercanamente relacionadas, las mutaciones usualmente conservan la función pero varían en la especificidad. Hay reportados varios ejemplos en la literatura, la familia de las serinas proteinasas contienen un bolsillo de especificidad, una superficie hendida complementaria en forma y distribución de carga a la cadena lateral adyacente al enlace escindible. Las mutaciones presentan una tendencia a dejar la conformación del esqueleto de carbono del bolsillo sin cambio, pero afectan la forma y carga de su revestimiento. Alterando la especificidad. Otro ejemplo lo tenemos en los pigmentos visuales, los cuales

cambian la sensibilidad espectral por mutaciones puntuales en residuos que interactúan con el cromóforo (Chothia *et al*., 2003)

Tales familias de enzimas ilustran el tipo de características estructurales que cambian y aquellos que permanecen iguales. En algunos casos, los átomos catalíticos ocupan la misma posición en el espacio molecular, aunque los residuos que los presentan están localizados en diferentes puntos en el plegamiento (Chothia *et al*., 2003; Akashi, 2012).

Existen varias familias de enzimas que muestran un mayor grado de divergencia. La superfamilia de endonucleasas apurínica/apirimidínica, un gran y diverso grupo de fosfoesterasas, incluye miembros que cortan DNA y RNA, además de fosfatasas lipídicas. Inclusive los residuos catalíticos varían entre las diferentes subfamilias de este grupo. Por ejemplo, una histidina esencial para la función reparadora de DNA de la enzima DNasa I no se encuentra conservada en la exonucleasa III. Contrariamente, muchas funciones son provistas por proteínas no relacionadas. La quimiotripsina y la subtilisina han producido el mismo mecanismo catalítico para la proteólisis mediante evolución convergente. Habiéndose observado una triada catalítica similar en un anticuerpo con actividad proteolítica, dicha triada catalítica aparece también en las lipasas.

Reclutamiento.

Una pregunta común en el área de evolución de proteínas es ¿Cuánto debe cambiar una proteína su secuencia antes de que cambie su función? Esta respuesta se complica por el hecho de que hay proteínas con múltiples funciones, pese a ser idénticas en secuencia, tal es el caso de las “moonlighting proteins” (Huberts *et al*., 2010).

1. Las proteínas del cristalino en patos son idénticas en secuencia a la lactato deshidrogenasa y la enolasa en otros tejidos, a pesar de no presentar el sustrato en el ojo. Ellas han sido reclutadas para proveer una función estructural y óptica. Existen otras proteínas del cristalino del ojo de las aves que son idénticas o similares a enzimas. En algunos casos residuos esenciales para la catálisis han

mutado, probando que la función de estas proteínas en el ojo no es enzimática. En estas especies, la coexistencia de enzimas mutadas e inactivas en el ojo y enzimas activas en otros tejidos implica que el gen correspondiente debe haber sido duplicado (Piatigorsky *et al*.1994; Piatigorsky *et al*., 1989).

2. Algunas proteínas interactúan con otras para producir oligómeros con diferentes funciones. En *E. coli*, una proteína que funciona por sí misma como la lipoato deshidrogenasa es también una subunidad esencial de la piruvato deshidrogenasa, de la 2-oxoglutarato deshidrogenasa y del complejo de corte de la glicina (Huberts *et al*., 2010).

3. Existen proteínas como DegP de la familia de las (HtrA) que funcionan como chaperonas a bajas temperaturas y como proteinasas a altas temperaturas. La lógica, aparentemente, es que bajo condiciones de estrés moderado ésta funciona rescatando proteínas mal plegadas; mientras que en condiciones de alto estrés cambia su función para reciclarlas (Sengupta *et al*., 2008).

4. La actividad de la fosfoglucosa isomerasa (neuroleucinas, factor de movilidad autocrino, mediador de diferenciación y maduración) depende de su localización. Estas proteínas funcionan como una enzima glucolítica en citoplasma, pero como un factor de crecimiento y citosina fuera de la célula (Sriram *et al*., 2005).

La divergencia y el reclutamiento son los extremos de un amplio espectro de cambios en secuencia y función. Aparte de casos de reclutamiento puro tal como el de las proteínas del cristalino del ojo de los patos o de la fosfoglucosa isomerasa, en las cuales una proteína adopta una nueva función sin cambio de secuencia en absoluto, hay ejemplos de cambios relativamente pequeños en secuencias correlacionados con muy pequeños cambios funcionales (lo cual la mayoría de la gente entendería como divergencia pura) y cambios relativamente pequeños en secuencias con cambios bastante grandes en función (lo cual la mayoría de la gente entendería como reclutamiento), pero también muchos casos en los que hay grandes cambios tanto en secuencia como en función.

Mezclado y combinado de dominios.

En esta categoría se incluyen los procesos de duplicación y oligomerización, e intercambio y fusión de dominios (Letunic *et al*., 2012; Goncearenco *et al*., 2015).

Muchas proteínas contienen ensambles en tándem de dominios que aparecen en diferentes contextos y orden en diferentes proteínas. Censos de genomas sugieren que muchas proteínas son multimodulares. De 4,401 genes en *E. coli*, 287 corresponden a proteínas que contienen dos, tres o cuatro módulos. Los patrones estructurales de 510 enzimas de *E. coli* involucradas en el metabolismo de pequeñas moléculas pueden ser representadas en su totalidad o en parte por 213 familias de dominios, mientras que de las 399 que pueden ser divididas dentro de dominios conocidos, 68% son proteínas de un solo dominio, 24% contienen dos dominios y 7% contienen tres dominios. Solamente cuatro de las 399 tuvieron cuatro, cinco o seis dominios. Se ha observado una preferencia marcada para la asociación de dominios pertenecientes a diferentes familias (Goncearenco *et al*., 2015; Letunic *et al*., 2012).

Las proteínas multidominios presentan problemas particulares a la hora de asignarles función en anotaciones del genoma, ya que los dominios pueden poseer función independiente, modular una u otra función o actuar concertadamente para proveer una única función que puede depender de la composición de dominios e incluso del orden. Por otro lado, en algunos casos la presencia de un dominio en particular o combinación de dominios está asociada con una función específica. Por ejemplo, el dominio de unión a NAD aparece casi exclusivamente en deshidrogenasas (Goncearenco *et al*., 2015).

2.4.1 Evolución proteica a nivel de ensamble de dominios.

Comparaciones hechas entre secuencias y estructuras proteicas muestran que el dominio es una importante unidad de evolución proteica. Los distintos dominios aparecen en una variedad de proteínas en diferentes combinaciones. Por lo tanto,

de una lista relativamente pequeña de dominios, la evolución puede ensamblar un gran número de proteínas completas (Goncearenco *et al*., 2015).

Estudios basados sobre la estructura de proteínas conocidas indican que es posible definir, aproximadamente 1000 superfamilias de dominios. De los aproximadamente 23 000 genes humanos, casi dos tercios contienen dominios conocidos. Las aproximadamente 1000 superfamilias de dominios contabilizan para aproximadamente 30 000 asociaciones en genes humanos (Goncearenco *et al*., 2015).

La población de dominios codificados por genes conocidos está distribuida de manera desigual. Nueve superfamilias de dominios contabilizan el 20% de los dominios emparejados en genes humanos. Algo similar se observa en otros eucariontes como Fugu, *D. melanogaster* y *C. elegans*.

La distribución de dominios depende de la clase funcional de la proteína. De igual forma, el número de proteínas en una clase funcional dada aumenta exponencialmente con el tamaño del genoma.

2.5 Las secuencias proteicas cambian debido a mutaciones neutrales.

Charles Darwin propuso que las especies evolucionan por selección natural. La idea de la selección positiva es aquella en la cual un fenotipo determinado es favorecido sobre otros, causando que la frecuencia alélica cambie a través del tiempo en dirección de aquel fenotipo. Bajo este tipo de selección natural, los alelos que confieren alguna ventaja son favorecidos como consecuencia de diferencias en la supervivencia y la reproducción entre los distintos fenotipos, llegando a ser dominantes en la población. De esta manera, sus fenotipos son retenidos. El que se favorezca o no un alelo es independiente de si es dominante o recesivo.

Contrariamente, a la selección positiva tenemos a la selección negativa, la cual elimina aquellos individuos con rasgos deletéreos. De acuerdo al Darwinismo, los

cambios en el fenotipo en una población son conducidos únicamente por selección positiva. Mientras que la visión neo-Darwinista de la evolución, o seleccionismo, mantiene que la selección natural es la principal fuerza responsable de los cambios a nivel molecular, es decir, en secuencias proteicas. La base del neodarwinismo es que existe una variabilidad lo suficientemente alta en cualquier población sobre la cual la selección positiva puede actuar. Las mutaciones son la fuente de variabilidad genética, pero cuales mutaciones son retenidas y pasadas a la descendencia es determinado por la selección positiva.

Esto último se puso a prueba por estudios realizados en 1980, cuando dos genes de la hemoglobina fueron secuenciados. A pesar de que ambos genes codificaban para el mismo producto, sus secuencias de nucleótidos diferían en un pequeño porcentaje. Además, se ha observado para ciertas proteínas que el número de reemplazamiento de aminoácidos aumenta de manera lineal con el paso del tiempo, a partir de que dos especies divergieron. Se ha observado de igual manera, que las sustituciones ocurren predominantemente en regiones de la proteína menos comprometidas con la estructura o la función, mientras que aquellas regiones trascendentales para una proteína difícilmente presentan alguna sustitución.

Por ejemplo, los fibrinopéptidos tienen una tasa de cambio mucho más alta que el citocromo c o las hemoglobinas, porque la mayor parte de la secuencia de los fibrinopéptidos no es importante para la función. Mientras que los residuos en el citocromo c y la hemoglobina que son más importantes para la función, aquellos que interactúan directamente con los grupos hemo, cambian mucho más lento que los residuos en la superficie de la proteína. Ninguna de estas observaciones podría ser explicada si la selección positiva fuera el principal medio de evolución en las secuencias (Kimura, 1981; Kimura, 1968).

2.5.1 Mutaciones neutrales son aquellas que no afectan la función

Una explicación mucho más simple fue propuesta por Kimura en 1968 y King y Jukes en 1969 conocida como la teoría neutral de la evolución molecular. Dicha teoría sostiene que una pequeña minoría de mutaciones en el DNA o secuencias proteicas son ventajosas y por lo tanto son fijadas por selección natural, mientras que otras son desventajosas y son eliminadas por selección natural purificadora, por último la gran mayoría de mutaciones que son fijadas son neutrales con respecto a la adecuación, siendo fijadas por medio de la deriva génica. En el caso de secuencias proteicas, tales sustituciones de aminoácidos no afectan de manera significativa la estructura y función de una proteína. De acuerdo a esta teoría, la mayoría de las variaciones genéticas a nivel molecular son selectivamente neutrales y carecen de significado adaptativo. Sin embargo, tales sustituciones ocurren a una tasa constante, de tal manera que el grado de diferencia de secuencias entre especies puede servir como un reloj molecular, dándonos la herramienta para poder determinar el tiempo de divergencia entre especies (Guo *et al*., 2004).

Debido a que la selección natural actúa principalmente en contra de mutaciones deletéreas, las mutaciones neutrales son toleradas, y con el paso del tiempo son fijadas en una población. Dado que las distintas regiones de una proteína pueden resistir un mayor o menor número de mutaciones, tenemos que la tasa de mutación de un aminoácido provee una medida de la importancia de este residuo para la estructura o función de una proteína. Algunas regiones presentan un menor número de sustituciones que otras, debido a que dichas regiones suelen ser esenciales para la estructura o función, como pueden ser; el sitio activo, los sitios de unión e inclusive aquellas superficies encargadas en el reconocimiento entre proteínas. Por lo tanto, para la teoría neutral la mayoría de los cambios que ocurren durante la evolución son aquellos que no afectan la función ni la estructura. (Kimura, 1981; Kimura, 1968).

Aquellas proteínas que presenten una baja tolerancia a mutaciones evolucionarán por ende de una manera lenta comparada con aquellas proteínas que toleren un mayor número de mutaciones en su secuencia. Quizás el mejor ejemplo conocido es el del citocromo c. La explicación del porque la baja tolerancia a mutaciones por parte de esta proteína es sencilla, debido a que el citocromo c toma parte en la transferencia de electrones, en la cadena respiratoria. Tal proteína debe interactuar con una variedad de otras proteínas, en el paso de los electrones a lo largo de la cadena de transporte de electrones, por lo que la cantidad de aminoácidos involucrados en interacciones con otras proteínas es elevado, esto si se compara con el tamaño relativamente pequeño del citocromo c, dejando poco espacio a mutaciones sin afectar las interacciones con otras proteínas. Además, los residuos cercanos al sitio de unión hemo están también conservados porque éste es donde la función del citocromo c reside. Por lo tanto, el número de mutaciones neutrales posibles es pequeño y el citocromo c ha cambiado muy lentamente en el curso de la evolución (Kimura, 1968; Margoliash, 1963).

2.5.2 Las mutaciones más aceptadas son neutrales.

De acuerdo a la teoría neutral de la evolución molecular, la mayoría de los cambios evolutivos y de la variación dentro y entre especies no es causada por la selección natural, sino por la deriva génica que se encarga de fijar mutaciones neutrales. Según esta teoría la mayoría de las mutaciones que ocurren son eliminadas por selección natural, por tratarse de mutaciones deletéreas. Mientras que el resto de las mutaciones, aquellas que no presentan una desventaja para el organismo, son en su mayoría neutrales y no beneficiosas como postula el seleccionismo. Esto explica por qué algunas proteínas mutan mucho más rápido que otras. Ya que para que una mutación sea fijada o no el único requisito es que no debe afectar de manera considerable la estructura y función de la proteína. Debido a esto las secuencias de proteínas cuyas funciones son menos tolerables a mutaciones presentan una tasa de mutación más lenta que aquellas secuencias que son más tolerables (Duret, 2008).

En conclusión no importa que tan rápido el DNA mute. Las proteínas difieren una de la otra en cuánto la secuencia puede ser cambiada sin producir efectos notables, es decir, sin afectar la función (Duret, 2008).

De acuerdo a la teoría neutral de la evolución, el proceso por el cual la mayoría de las mutaciones son fijadas es la deriva génica. Por ende, en una población finita, un gran número de mutaciones serán fijadas de manera aleatoria, independientemente de si confieran ventaja alguna a los individuos que las portan. Esto claramente contrasta con el seleccionismo, el cual propone que para que una mutación se fije, ésta debe conferir alguna ventaja al individuo (Kimura, 1968; Hughes, 2007; Chun *et al*., 2011).

La selección neutral no niega la existencia de selección positiva, sin embargo, el número de mutaciones que resultan de ésta es mucho más pequeño que el número de mutaciones neutrales. De hecho los expertos en el tema concuerdan con que ambas teorías son compatibles (Kimura, 1968; Hughes, 2007; Chun *et al*., 2011).

2.5.3 Cambios importantes en las proteínas resultan de la selección positiva.

Un tema central en la polémica entre la teoría neutral de la evolución y la teoría de la selección natural tiene que ver con la proporción de sustituciones que resultan por efecto de la selección positiva. Claro está que este tipo de selección solo puede llevar a la fijación a aquellas mutaciones que son benéficas para el individuo. Y se espera que la frecuencia de aparición de mutaciones ventajosas sea baja en tanto la función del gen permanezca sin cambio (Nei *et al*., 1984).

Pese a que la teoría neutral de la evolución postula que la mayoría de las mutaciones son neutrales, es cierto también que dicha teoría no niega la selección natural, de hecho algunas de las más importantes mutaciones en la adaptación son el resultado de selección positiva. Teniendo en el ambiente, la principal fuerza o limitante que favorece un cambio. Por ejemplo, si hay una fuerte limitante

ambiental que requiera un rasgo determinado, ningún cambio puede ocurrir sobre este rasgo. Contrariamente, un cambio en el ambiente, tal como la modificación del hábitat, o la aparición de nuevos químicos en éste, conducen hacia una rápida evolución. A continuación se citan algunos ejemplos (Nei *et al.,* 1984; Almeida 2016).

Uno de los primeros casos donde se pudo presentar evidencia rotunda de la existencia de selección positiva, lo constituye el ejemplo del gen que codifica la lisozima en rumiantes y en los langúridos, estos últimos pertenecientes al orden de los primates. En estos dos grupos de mamíferos la enzima es secretada en el estómago anterior donde cumple la función de digerir la pared celular bacteriana, lo cual permite utilizar la celulosa degradada por las bacterias, que habitan el tracto digestivo de tales animales.

Sin embargo, en otros grupos de mamíferos esta enzima no se secreta en el estómago. Las secuencias proteicas de estas enzimas contienen algunos aminoácidos que están presentes solamente en estos dos grupos de mamíferos. Otros grupos de primates presentan secuencias que carecen de dichos aminoácidos. Sin duda no podemos atribuir la similitud que se observa para esta proteína entre langúridos y rumiantes a ancestría común (esto es son similares porque el ancestro común entre ambos ya poseía dichos aminoácidos), puesto que esta explicación implicaría asumir que los langúridos son más próximos a los rumiantes que al resto de los primates. La explicación más razonable es que en ambos linajes estos aminoácidos se adquirieron independientemente, por lo que la lisozima representaría un ejemplo de convergencia a nivel molecular. Podríamos argumentar que esta convergencia se debe a sustituciones neutras que se adquirieron al azar y en paralelo en ambos grupos de mamíferos, sin embargo esta alternativa es poco probable si tomamos en cuenta que las mismas involucran por lo menos siete aminoácidos, cada uno de los cuales tiene una probabilidad de (1/19) de ser convergente por azar. La probabilidad conjunto para las siete sustituciones convergentes es menor de 10^{-6}. Mucho más razonable, en cambio, es la explicación que propone que dichos cambios en los aminoácidos de

manera paralela fueron el resultado de una misma respuesta adaptativa, es decir darle capacidad a la lisozima de funcionar correctamente a un pH bajo, como el que se encuentra en el estómago anterior de estos animales. La lisozima humana, por ejemplo, funciona en forma muy ineficiente en condiciones de acidez. Además se ha podido determinar que algunas de estas sustituciones de aminoácidos confieren mejor desempeño a la enzima en condiciones ácidas. Este ejemplo en particular representa un ejemplo de respuesta adaptativa a un cambio en el nicho ecológico (Nei *et al*., 1984; Miyata *et al*., 1980; Almeida, 2016).

Otro ejemplo lo tenemos con la hemoglobina de los cocodrilos, la cual ha sufrido varias mutaciones provocando un cambio funcional, dicha hemoglobina en lugar de unir fosfatos orgánicos como bifosfoglicerato, Cl^- y CO_2; une bicarbonato, característico del ambiente ácido de los tejidos del cocodrilo. De tal manera que la hemoglobina de estos reptiles es capaz de liberar una mayor cantidad de oxígeno, provocando que los cocodrilos hagan un uso más eficiente del oxígeno cargado en sus pulmones. Tal adaptación se debe al cambio de solamente 5 aminoácidos de un total de 123 que presenta la molécula de hemoglobina. Cabe mencionar que han sido ubicados un mayor número de cambios en la secuencia proteica de la hemoglobina, los cuales no juegan un papel aparente en la función de esta molécula, por lo que se consideran neutrales. De manera análoga los camélidos andinos, cuyo hábitat de alta montaña ha hecho que estos animales se adapten a la baja presión de oxígeno, presentan unas pocas sustituciones de aminoácidos que provocan un aumento de la afinidad de la hemoglobina por el oxígeno (Nei *et al*., 1984; Miyata *et al*., 1980; Almeida, 2016).

Otro ejemplo son las opsinas de los conos en la retina de seres humanos, las cuales son las responsables de la visión a color y son el resultado de una duplicación génica. El fenotipo original era verde, a partir del cual se originó el rojo por un proceso de duplicación y una serie de mutaciones sobre una de las copias. Existen 15 aminoácidos diferentes entre ambas proteínas, pero solamente dos de estas son importantes para la diferencia en función entre las dos opsinas. En peces las opsinas se han adaptado a la luz tenue de aguas profundas por la

misma sustitución de aminoácidos en al menos ocho ocasiones distintas. Dado el hecho de que varias de estas mutaciones ocurrieron al mismo tiempo, es un fuerte indicio de selección positiva (Almeida, 2016).

2.5.4 La estructura proteica codifica un reloj molecular.

Ahora se sabe que proteínas distintas tal como la hemoglobina y el citocromo c, evolucionan a diferentes tasas. Asimismo, se sabe que las proteínas homólogas parecen mutar a una tasa constante, independientemente de si una determinada proteína se encuentra en uno u otro organismo. Es decir, la tasa de cambio de secuencia es constante en tiempo real para cada proteína. Esto fue descubierto por Zuckerkandl y Pauling en 1965, cuando encontraron que la cantidad de diferencias entre aminoácidos de las cadenas de hemoglobina de los distintos linajes correlacionaba con la tasa evolutiva de divergencia propuesta por la evidencia fósil. A partir de su estudio propusieron que “por lo tanto es posible que exista un reloj molecular evolutivo” el cual determina la tasa de cambio de una secuencia proteica.

Dado que las enzimas encargadas de la replicación del DNA son muy similares entre especies, se asumió en un principio que la tasa de error de replicación del material genético es constante, no solo a lo largo del tiempo, sino en todas las especies y en cualquier parte del genoma que se quiera comparar. Ahora se sabe que una tasa de mutación constante en el DNA no es la responsable de la tasa de mutación constante observada en proteínas, ya que de ser así todas las proteínas de un mismo organismo habrían evolucionado a la misma velocidad. Por lo tanto, el reloj evolutivo está determinado por la propia proteína y no por su DNA. En otras palabras, la secuencia proteica cambia principalmente a través de la acumulación de mutaciones neutrales (Kimura, 1968; Zuckerkandl *et al*., 1962; Zuckerkandl *et al*., 1965).

El concepto de reloj molecular revolucionó el estudio de la evolución de proteínas. Si los relojes son construidos a partir de proteínas que presentan una tasa de

mutación muy baja, es decir, que los cambios en la secuencia de una proteína son fijados muy lentamente y si suponemos una tasa constante de cambio de estas secuencias en el tiempo evolutivo, podemos entonces extrapolar en el tiempo y determinar el punto en el que ocurrieron eventos evolutivos que no son accesibles por medio del registro fósil, es decir, mayores a 2.5 millones de años. Contrariamente si la tasa de mutación es alta y constante en una proteína en cierto grupo de linajes, podemos con ayuda de esta última, establecer eventos filogenéticos no documentados por el registro fósil, como son la divergencia de taxa vivos, siendo este un acontecimiento demasiado reciente para ser captado por el registro fósil (Kimura, 1968; Morgan, 1998; Zuckerkandl *et al*., 1962; Zuckerkandl *et al*., 1965).

3 Actividad Alcohol deshidrogenasa en animales

La actividad de alcohol deshidrogenasa (ADH) se encuentra ampliamente distribuida en los tres dominios de la vida. Encontrándose representada por tres distintas superfamilias en animales, que se originaron independientemente a lo largo de la historia evolutiva del linaje de los animales. Por lo tanto, no presentan homología, además de presentar mecanismos de reacción distintos, así como una distinta estructura. La primera de estas superfamilias corresponde a las ADH dependientes de zinc o de tipo I, siendo el grupo de enzimas más estudiado y mejor comprendido a la fecha; posteriormente tenemos a las ADH de cadena corta o de tipo II, de las cuales su actividad de alcohol deshidrogenasa se encuentra restringida a organismos del genero Drosophila; por último tenemos a las ADH dependientes de hierro o de tipo III de las cuales ha habido muy pocos estudios, siendo de las tres superfamilias de ADH presentes en animales de la que se conoce menos en todos los aspectos. El objetivo principal del presente trabajo, fue realizar el primer análisis filogenético sobre esta última superfamilia. A continuación se dará una mayor explicación de cada una de las superfamilias de ADH de animales (Hernández *et al*., 2011; Riveros *et al*., 2003; Jörnvall *et al*., 2015).

3.1 ADH dependientes de Zinc

Este grupo de enzimas fue descubierto hace más de 100 años cuando Federico Battelli y Lina Stern, probaron por vez primera una preparación soluble de alcohol deshidrogenasa obtenida de un macerado de hígado de caballo. Posteriormente Beng Andersson demostró que este tipo de enzimas solo presentan su actividad como alcohol deshidrogenasas en presencia de NAD^+. Y fue en 1937 cuando se purificó y cristalizó por primera vez una enzima de esta superfamilia, la cual pertenecía a, *Saccharomyces cerevisiae*, aunque en 1948 Bonnichsen y Wassen lograron cristalizar la ADH de hígado de caballo (Jörnvall *et al*., 1970).

Desde el punto de vista bioquímico las alcohol deshidrogenasas de tipo I, con número EC 1.1.1.1, son un grupo de deshidrogenasas presentes en una gran variedad de organismos, cuya función es facilitar la inter-conversión entre alcoholes y aldehídos o cetonas con la reducción de nicotinamida adenina dinucleótido (NAD^+ a NADH). En animales, incluyendo al ser humano, juegan un importante papel en la oxidación de alcoholes que de otra manera resultarían tóxicos al organismo, además de participar en la generación de aldehídos, cetonas o grupos alcohol durante la biosíntesis de una gran variedad de metabolitos. En levaduras, plantas y muchas bacterias, algunas enzimas pertenecientes a este grupo catalizan la reacción opuesta como parte de la fermentación para asegurar un constante suministro de NAD^+ permitiendo que la glucólisis pueda continuar. Las alcohol deshidrogenasas en animales comprende un grupo de varias isoenzimas que pueden catalizar la oxidación de alcoholes a aldehídos y cetonas (Jörnvall H. *et al*., 1970).

En cuanto al origen de esta superfamilia, existe evidencia que surge de la comparación de secuencias en múltiples organismos que muestran una formaldehido deshidrogenasa dependiente de glutatión, idéntica a la alcohol deshidrogenasa de la clase III (ADH-3/ADH5), por lo que se propone que las enzimas de tal clase pudieron actuar como la enzima ancestral de la familia entera de las ADH-Zn. En los primeros organismos, un método eficaz para la eliminación de formaldehído tanto endógeno como exógeno era importante, habiéndose

conservado esta capacidad en las ADH-3 a través del tiempo. Posteriormente, mediante un proceso de duplicación génica seguido de una serie de mutaciones, se habrían originado las enzimas ADH presentes en las restantes clases. (Sanghani *et al*., 2002; Gutheil, *et al*., 1992; Danielsson *et al*., 1992).

Por otro lado, la habilidad para producir etanol a partir del azúcar, puede haberse originado inicialmente en levaduras. Aunque esta característica no es adaptativa desde el punto de vista energético, si lo es desde el punto de vista en que altas concentraciones de alcohol pueden ser tóxicas para otros organismos, por lo que las levaduras podrían eliminar la competencia por el sustrato, en los frutos en descomposición. Debido a que los frutos podridos pueden contener una relativa alta concentración de etanol, los animales que consumen dichos frutos necesitan un sistema para metabolizar ese etanol exógeno, por lo que se pensó que ésta sería la explicación de haberse conservado la actividad oxidante de etanol en las ADH en otras especies distintas a la levadura. Sin embrago, ahora se sabe que las enzimas de la clase III (ADH-3) tienen un papel muy importante en la señalización del óxido nítrico, mientras que las enzimas de las restantes clases jugarían una diversidad de funciones en el metabolismo de una variedad de metabolitos internos (Sanghani *et al*., 2002; Godoy 2006 *et al*., 2006; Jörnvall *et al*., 1970).

Vale la pena destacar que en seres humanos, la secuencia del gen ADH1B, que codifica para una ADH, muestra dos variantes, en una de las cuales se observa un polimorfismo de un sólo nucleótido (SNP), el cual conduce al cambio de un residuo de histidina o uno de arginina en la enzima que cataliza la conversión de etanol a acetaldehído. En la variante que presenta el residuo de histidina la enzima es mucho más efectiva en la conversión de etanol a acetaldehído. Debido a que la enzima responsable para la conversión de acetaldehído a acetato, permanece sin cambio, la cual conduce a una tasa diferencial de la catálisis del sustrato causando una acumulación de acetaldehído tóxico, provocando daño celular. En seres humanos, varios haplotipos derivados de esta mutación se encuentran más concentrados en las regiones cercanas al este de China, una región cuyos

pobladores presentan una baja tolerancia y dependencia al alcohol (Whitfield, 1994).

Adicionalmente, se realizó un estudio con el fin de encontrar una correlación entre la distribución alélica y el alcoholismo, y los resultados sugieren que la distribución alélica surge en la misma región junto al cultivo de arroz hace entre 12 000 y 6 000 años. En dichas regiones donde se cultivaba el arroz, dicho grano también era fermentado para producir bebidas alcohólicas. Los resultados de un aumento de la disponibilidad de alcohol llevó al alcoholismo y al abuso a aquellos capaces de adquirirlo, lo que resultó en una menor adecuación reproductiva (reproductive fitness). Mientras que las personas con la variable del alelo tienen poca tolerancia al alcohol, lo que disminuye la oportunidad de dependencia y abuso. Según esta hipótesis postula que aquellos individuos con la enzima histidina-variante eran lo suficientemente sensibles a los efectos del alcohol, provocando un éxito reproductivo diferencial de los alelos correspondientes, que se heredaron a través de las generaciones (Whitfield, 1994; Peng *et al*., 2010).

De acuerdo a esta hipótesis la selección positiva actuaría para seleccionar en contra de la forma perjudicial de la enzima (variante con arg) debido al bajo éxito evolutivo de los individuos portadores del alelo. El resultado sería una mayor frecuencia del alelo responsable de la enzima con la variante de his en regiones que habían estado bajo presión selectiva por más tiempo. La distribución y frecuencia de la variante con his, a su vez, sigue la dispersión del cultivo del arroz hacia regiones del interior de Asia, con una mayor frecuencia de la variante de his en regiones que han cultivado el arroz por más tiempo. La distribución geográfica de los alelos parece por lo tanto ser el resultado de la selección natural en contra de individuos con un menor éxito reproductivo, es decir aquellos que llevaban el alelo con la variante arg y que por lo tanto eran más susceptibles al alcoholismo. (Whitfield, 1994; Peng *et al*., 2010).

Sin embrago, lo que no comentan los autores al respecto del párrafo anterior, es que la esperanza de vida en la población general fue de 30 años hasta el siglo

XIX, y que las bebidas alcohólicas eran un producto caro como para que la población general pudiera consumirlo en grandes cantidades. El alcoholismo como problema social surge hasta el desarrollo de los aparatos de destilación continua que permitieron la producción barata de bebidas destiladas.

En seres humanos existen múltiples formas de ADHs las cuales son codificadas por al menos siete diferentes genes. Existen 5 clases de alcohol deshidrogenasas, pero la forma hepática que es usada principalmente en humanos pertenece a la clase I. Dicha clase consiste de subunidades α, β, y γ que son codificadas por los genes ADH1A, ADH1B y ADH1C. Tales enzimas están presentes en altos niveles en el hígado y la mucosa del estómago. Catalizando la oxidación del etanol a acetaldehído (Sultatos *et al*., 2004; Farrés *et al*., 1994).

$$CH_3CH_2OH + NAD^+ \rightarrow CH_3CHO + NADH + H^+$$

Esto permite el consumo de bebidas alcohólicas, aunque su propósito, según algunos autores es probablemente la descomposición del alcohol contenido de forma natural en los alimentos o producido por las bacterias en el tracto digestivo (Kovacs *et al*., 2011).

Por otro lado, hay quienes plantean que el propósito podría ser el metabolismo de alcoholes endógenos, como la vitamina A (retinol), el cual se genera a partir de la hidrolisis de éster retinílico y de la reducción de retinal (retinaldehído), aunque la función aquí puede ser principalmente la eliminación de niveles tóxicos de retinol. Además, la enzima podría cumplir una variedad de papeles en el metabolismo de muchas moléculas endógenas, lo cual tiene como fundamento que las diferentes enzimas de las ADH dependientes de zinc presentan mayor afinidad y eficiencia catalítica para dichas moléculas que para el etanol (Duester, 2008; Hellgren *et al*., 2007).

3.2 ADH de cadena corta

Esta familia de alcohol deshidrogenasa fue descrita inicialmente en organismos del genero Drosophila, y se encuentran exclusivamente en el Orden Díptera de los

insectos. De acuerdo con los análisis filogenéticos, estas enzimas derivan de las prostaglandinas deshidrogenasas (PGDH), las cuales están presentes en todos los animales. Siendo precisamente las PGDH a partir de las cuales, por procesos de duplicación se originaron las ADH de cadena corta y otras enzimas relacionadas. Cabe resaltar que el género Drosophila, el cual presenta la enzima funcional para oxidar etanol, tuvo su origen hace 100 millones de años, esto es, después de que las Angiospermas productoras de frutos que pueden ser fermentados, tuvieran su origen. Además, se ha demostrado que la resistencia al alcohol esta positivamente correlacionada con la actividad de esta enzima en dichos organismos, por otro lado, la expresión de este gen sí está regulada por su sustrato, en este caso el etanol. De hecho, dada la gran diversidad de organismos dentro del género Drosophila, se ha propuesto que esta enzima ha conferido una ventaja sobre los demás organismos para ocupar un nicho ecológico antes vacío, los frutos en descomposición, explicando así su éxito evolutivo (Hernández *et al*., 2011).

3.3 ADH dependientes de hierro

Se trata de la superfamilia de ADHs menos estudiada de las tres encontradas en animales. Fueron inicialmente descritas en procariontes, donde presentaban una clara función como enzimas oxidantes de etanol. Posteriormente fueron encontradas enzimas con una variedad de funciones como lactaldehído reductasa, glicerol deshidrogenasa, butanol deshidrogenasa, propanodiol deshidrogenasa y otras que utilizan etanolamina. Todas estas enzimas fueron identificadas como miembros de las ADH-Fe y están presentes en diversos procariontes. Cabe mencionar que desde inicios del año 2000 a la fecha, han sido encontradas prácticamente en todos los animales estudiados, entre ellos el ser humano. Se trata de un gen de copia única y dado que se encuentra presente en todos los animales, tenemos que su origen es anterior a la exposición natural al etanol. Además, estudios inmunocitoquímicos, así como análisis de predicción en la ubicación de dichas proteínas con MITOPRED, WoLF PSORT y PredSL, sugieren que se trata de proteínas mitocondriales, al menos en animales. Pese a esto

último, dichas enzimas siguen presentando ambos motivos; un motivo de unión a NAD(P), caracterizado por un plegamiento tipo Rossmann y el motivo de unión a Fe^{+}. La falta del tercer residuo de histidina, hace que otros iones divalentes puedan interactuar con dicho motivo de unión a cofactor, teniéndose un claro ejemplo en la unión del Zn en la ADH4 de *Saccharomyces cerevisiae* (Carter y Gibson 2005; Ji-Hyun *et al*., 2011; Kim *et al*., 2007; Guda *et al*., 2004; Horton *et al*., 2007; Petsalaki *et al*., 2006).

Se ha visto que la exposición a etanol no modifica los niveles de expresión del gen adhfe1. Se ha demostrado que la expresión de dicho gen se encuentra aumentada en caballos de carreras, comparados con su contraparte doméstica; en ambos casos la enzima cataliza la conversión del ácido γ-hidroxibutírico a succinato semialdehído en una reacción acoplada con la reducción de α-cetoglutarato. También se ha demostrado que ciertas sales de ácido succínico elevan los niveles de hormona de crecimiento, siendo esta una manera plausible de explicar el elevado tono muscular de los animales de competencia. Además, se ha demostrado que dicha enzima puede presentar actividad catalítica sin tener que añadir $NAD(P)^{+}$, el cual generalmente actúa como el aceptor de electrones en el metabolismo del etanol, por lo que difícilmente el etanol puede haber ejercido presión de selección sobre esta familia de ADH (Ji-Hyun *et al*., 2011).

Cabe mencionar que la función del ácido γ-hidroxibutírico dentro de las células no es clara aún, aunque en estudios recientes se ha observado su formación en el cerebro de los vertebrados como producto del metabolismo del neurotransmisor GABA (Carter y Gibson, 2005).

En el presente proyecto pretendemos contribuir a la comprensión de los mecanismos que dirigen la evolución de proteínas y más específicamente, el proceso de enzimogénesis, ya que mediante el análisis de las diferencias y similitudes a nivel de secuencias, y el establecimiento de un posible patrón filogenético en esta familia de alcohol deshidrogenasas, podremos esclarecer cuales son los cambios críticos que pudieron ocurrir para modificar la especificidad

de estas enzimas. Cabe mencionar que de la superfamilia de ADH-Fe no se ha publicado ningún análisis filogenético a la fecha.

4. Objetivos

Realizar un análisis filogenético de las alcohol deshidrogenasas dependientes de hierro que aporte información útil para entender mejor la historia evolutiva de esta familia de proteínas y la diversidad de funciones (actividades) desarrollada por estas enzimas en eucariontes.

1) Recuperar de las bases de datos de acceso público todas las secuencias de proteínas pertenecientes a la súper familia de alcohol deshidrogenasa dependiente de hierro.

2) Realizar el alineamiento estructural con aquellas secuencias de las cuales se conoce sus estructuras tridimensionales pertenecientes a la súper familia de alcohol deshidrogenasa dependiente de hierro.

3) Obtener el primer árbol filogenético para la superfamilia de alcohol deshidrogenasas dependientes de hierro por varios algoritmos.

5. Metodología

Se recuperaron de la base de datos del NCBI todas las secuencias proteicas no redundantes correspondientes a las alcohol deshidrogenasas activadas por Fe, se tomó como secuencia query la ADH-Fe perteneciente al ser humano, la búsqueda de homólogos se hizo mediante el algoritmo BLASTP, utilizando los valores predeterminados, los cuales incluyen la matriz de sustitución de BLOSUM62. Se recuperaron un total de 782 secuencias, sobre las cuales se inició un proceso de depuración, eliminando aquellas secuencias que no tuvieran al menos una longitud de 200 aminoácidos, recordando que las ADH-Fe tienen en promedio 470 aminoácidos, de igual forma se eliminaron secuencias repetidas, por tratarse de un gen de copia única, conservando aquellas secuencias que presentaran una

longitud de secuencia más cercana a los 470 aminoácidos (Altschul *et al*., 1990; Altschul *et al*., 1997; Larkin *et al*., 2007; Boratyn *et al*., 2013; Pundir *et al*., 2016).

Una vez depurado el número de secuencias se procedió a recuperar con ayuda de VAST (Vector Alignment Search Tool, http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml) en la página del NCBI todas aquellas secuencias de aminoácidos cuya estructura ha sido resuelta, generalmente por cristalografía de rayos X. Tomando como base aquellas 12 secuencias de las cuales se conoce su estructura tridimensional, se alinearon las restantes secuencias, lo que se conoce como alineamiento estructural. Lo anterior tiene como principio que a lo largo del proceso evolutivo de divergencia dentro de una superfamilia, se encuentra mejor conservada la estructura tridimensional de las proteínas que sus secuencias primarias. Posteriormente los alineamientos se generaron utilizando Clustal X (Thompson *et al*. 1997). y fueron corregidos a mano, con ayuda del programa BioEdit (Hall, 1999; Madej *et al*., 2014).

Para obtener la muestra representativa imparcial más pequeña de proteínas homólogas a las ADHFe, las secuencias proteicas fueron recuperadas de la base de datos de Pfam versión 29.0, basados sobre proteomas representativos con un umbral de co-pertenencia del 15% (RP15), de igual forma solo se recuperaron secuencias con más de 200 aa de longitud. Esto con la finalidad de hacer un análisis de correspondencia entre las familias proteicas descritas por la base de datos de dominios conservados del NCBI y la base de datos Pfam.

Una vez obtenido nuestro alineamiento se seleccionó el mejor modelo de sustitución con ayuda de MEGA 7, para posteriormente realizar los análisis filogenéticos con cuatro distintos algoritmos, los métodos basados en distancia; Vecino más cercano (NJ) y método de Mínima Evolución; así como por los métodos basados en caracteres que incluye; al método de Máxima Parsimonia y al Método de Máxima Verosimilitud con un valor de confianza de bootstrap para todos los algoritmos de 500 réplicas (Kumar *et al*., 2016). El modelo de sustitución de aminoácidos descrito por Le-Gascuel, usando una distribución gamma discreta con cinco categorías, fue el seleccionado como el mejor modelo de sustitución.

Adicionalmente, se hizo el análisis del contexto genómico para cada una de las secuencias pertenecientes a animales y hongos en nuestro listado de secuencias previamente depurado.

Por otro lado, con ayuda de softwares como MITOPRED, PSORT y PredSL, se predijo la ubicación de cada una de las secuencias pertenecientes al grupo de animales y de las secuencias de hongos cercanamente relacionadas (Kim *et al*., 2007; Guda *et al*., 2004; Horton *et al*., 2007).

De manera simultánea se recuperaron las secuencias de los genes de *adh-fe* en animales, además de hacerse el análisis filogenético correspondiente, previa determinación del mejor modelo de sustitución con ayuda de MEGA 7 (Kumar *et al*., 2016).

Igualmente las señales de selección positiva se analizaron en el servidor datamonkey http://www.datamonkey.org/. Después de que el análisis arrojó los codones sujetos a selección positiva, éstos fueron identificados en la secuencia proteica y graficados con ayuda del programa LOGOS.

Por último se utilizó el programa I-TASSER http://zhanglab.ccmb.med.umich.edu/ITASSER/. para predecir la estructura tridimensional de la ADH-Fe perteneciente a humano.

6. Resultados y discusión

Las alcohol deshidrogenasa dependientes de hierro o de tipo III, forman parte de la superfamilia de las ADH-Fe en las bases de datos de proteínas. Sin embargo como ya se mencionó en la sección de clasificación de proteínas, las distintas bases de datos emplean diferentes criterios para clasificar las secuencias proteicas dentro de diferentes familias e inclusive superfamilias de proteínas; dando por consiguiente que los límites entre familias de proteínas relacionadas no son homogéneos. La base de datos de dominios conservados del Centro Nacional para la Información Biotecnológica (NCBI) (Marchler-Bauer *et al*., 2012) agrupa a las ADH dependientes de hierro dentro de la superfamilia de proteínas DHQ Fe_ADH con número cd07766. Dicha superfamilia engloba a cuatro familias relacionadas: la primera de éstas es la familia dehidroquinato sintasa con número cd08169, la cual cataliza la conversión de 3-deoxi-D-arabino-heptulosonato-7-fosfato (DAHP) a dehidroquinato (DHQ) en el segundo paso de la vía del shikimato; la segunda familia es la de la glicerol 1 fosfato deshidrogenasa y proteínas relacionadas con número cd08549; la tercera comprende a la familia de glicerol deshidrogenasa con número cd08550; y por último a la familia de alcohol deshidrogenasa activada por Fe con número cd08551. Por otro lado, la base de datos Pfam (Finn *et al*., 2016) clasifica estas proteínas en tres diferentes familias proteicas: la primera es la familia de dehidroquinato sintasa con número PF01761; la segunda es la familia de alcohol deshidrogenasa activada por hierro con número PF00465; y por último a una segunda familia de alcohol deshidrogenasa activada por hierro con número PF13685.

Se realizó una prueba de correspondencia entre las familias de proteínas descritas en ambas bases de datos, por lo que se recuperaron todas las secuencias identificadas de la base de datos de dominios conservados del NCBI las cuales estaban distribuidas de la siguiente manera: 152 secuencias correspondían a la familia con número cd08169, 66 secuencias de la familia cd08549, 118 secuencias pertenecientes a la familia cd08550 y por último 538 secuencias agrupadas dentro de la familia cd08551. Todas estas secuencias fueron alineadas con muestras

representativas no sesgadas de secuencias de proteínas tomadas de las distintas familias de acuerdo a la clasificación de Pfam para las ADH activadas por hierro (tomando un umbral del 15% de co-pertenencia); de las cuales 518 secuencias pertenecían a la familia PF01761, 79 secuencias de la familia PF13685 y por último 1080 secuencias de la familia PF00465. Se obtuvo el dendrograma mostrado en la figura 1 (sección de resultados) mediante el método de máxima verosimilitud, en dicha figura se observa la correspondencia entre proteínas de las familias de la base de datos de dominios conservados del NCBI y las familias de Pfam. En dicha figura podemos observar que la familia de las dehidroquinato sintasa con número cd08169 comparte la misma rama que secuencias proteicas de la familia PF01761 de acuerdo a la base de datos de Pfam. Del mismo modo, la familia de glicerol 1 fosfato deshidrogenasa y proteínas relacionadas con número cd08549 están localizadas en la misma rama que las secuencias pertenecientes a la familia 2 de alcohol deshidrogenasa activadas por hierro de acuerdo a la base de datos de Pfam con número PF13685.

Por otro lado, la familia de alcohol deshidrogenasas activadas por hierro de la base de datos de Pfam con número PF00465 agrupa secuencias que pertenecen a dos familias de proteínas relacionadas reportado así por la base de datos de dominios conservados del NCBI; correspondientes a la familia de glicerol deshidrogenasa (cd08550) y la familia de alcohol deshidrogenasa activadas por hierro (cd08551). Debido a que ha sido reportado que las glicerol deshidrogenasas son metaloenzimas que contienen un átomo de zinc en lugar de hierro por Spencer y cols. 1989 y 1990, comprende una rama divergente con respecto a las otras alcohol deshidrogenasas que contienen hierro, como se aprecia en la figura 1 (sección de resultados), conservando solamente una de las tres histidinas involucradas en la unión a hierro, como se aprecia en el alineamiento múltiple de secuencias representado en la figura 2 (sección de resultados). Aquí es preciso aclarar que en el presente trabajo nos enfocamos únicamente a la familia de las alcohol deshidrogenasa dependientes de hierro, bajo la clasificación hecha por la base de datos de dominios conservados del NCBI (cd08551).

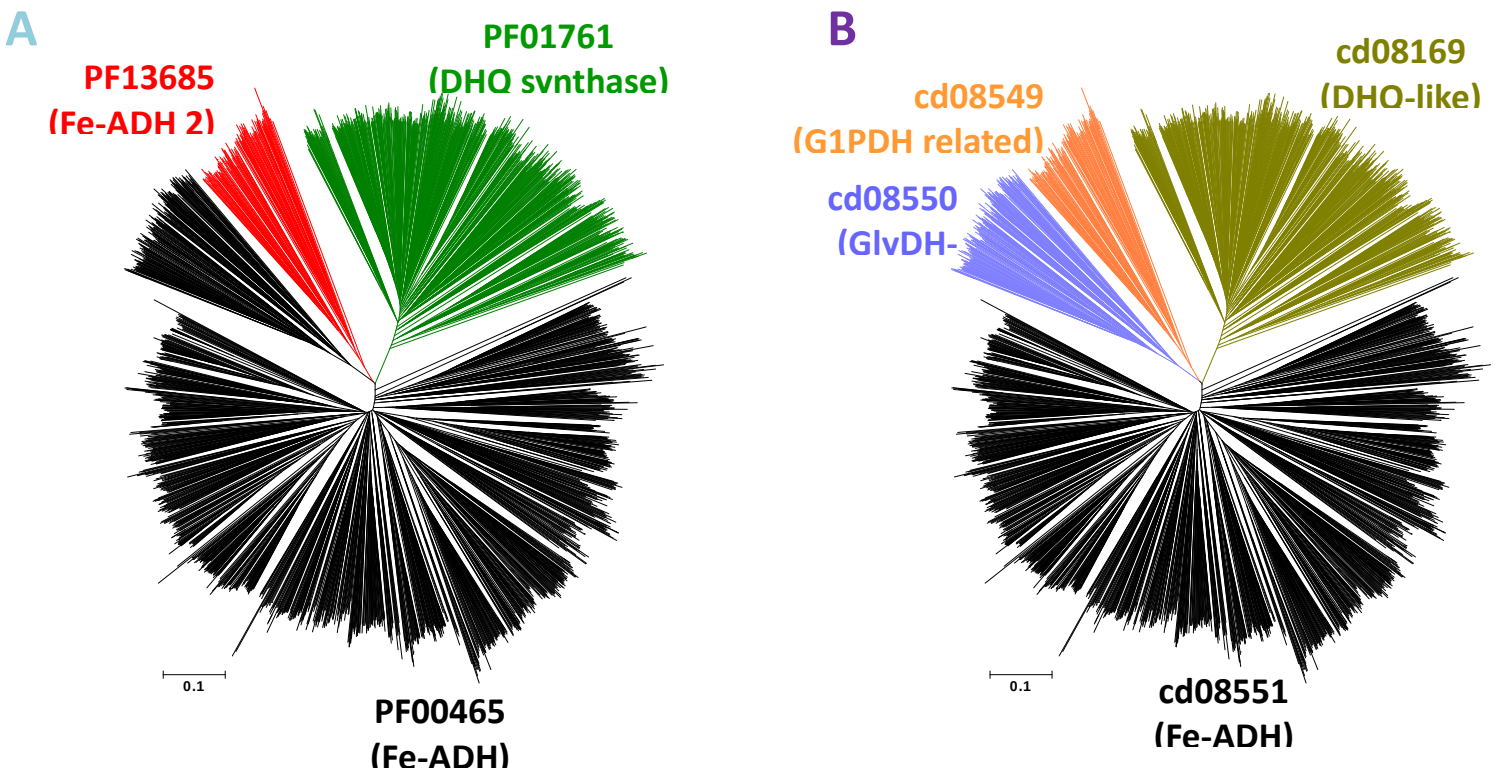
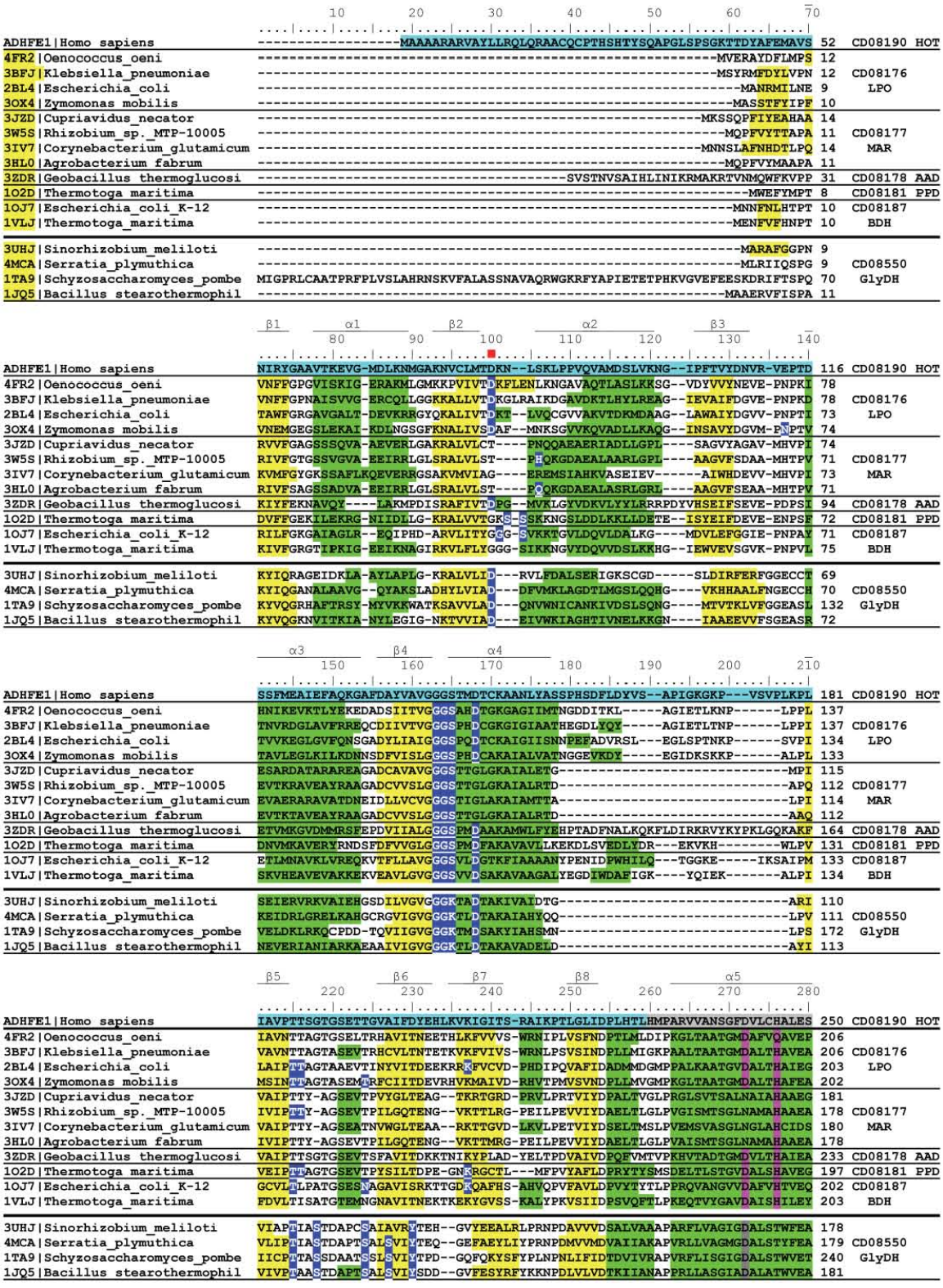


Figura 1. Se presenta un árbol sin enraizar con secuencias proteicas que poseen homología a las ADH-Fe. En dicho estudio se incorporaron 2459 secuencias no redundantes pertenecientes al Protein Data Bank, Swiss Prot database, NCBI's Conserved Domain Database, y Pfam database (usando la opción RP15 permitiendo así la mayor representación de proteínas divergentes). Las secuencias proteicas fueron agrupadas en familias de acuerdo a criterios de la **base de datos de Pfam (A)** o la **base de datos de dominios conservados de NCBI (B)**.



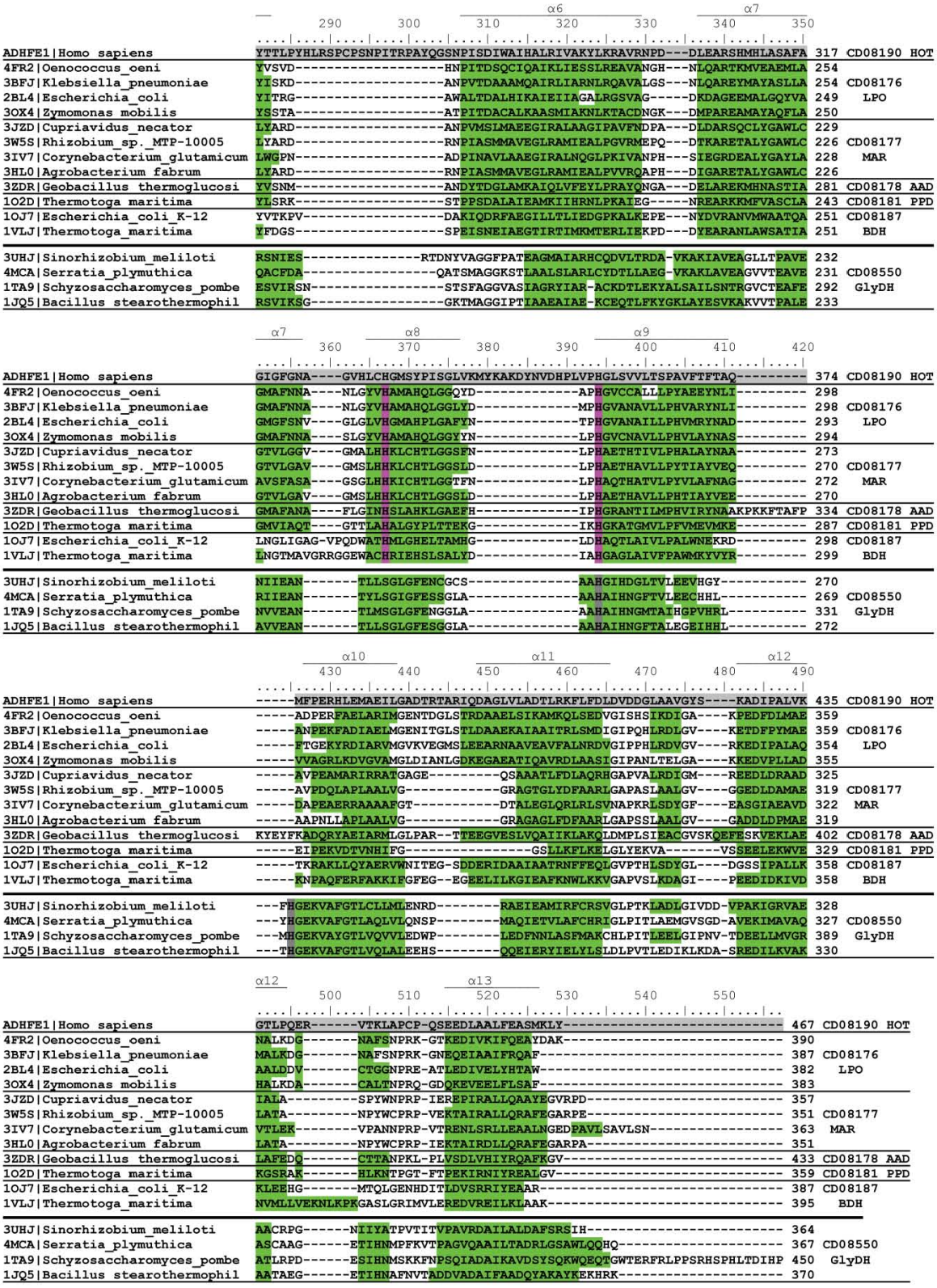


Figura 2. Se presenta el alineamiento estructural, que sirvió como base para el alineamiento múltiple de secuencias. Se ejemplifican todas las secuencias para las

cuales su estructura tridimensional ha sido resuelta y recuperada del PDB, de la familia de las ADH-Fe. Todas estas doce secuencias pertenecen a cinco diferentes subfamilias agrupadas en la familia ADH-Fe. De igual forma en la imagen previa, se incorporó la secuencia de ADH-Fe humano, con motivos de comparación, además de incluir 4 secuencias correspondientes a las glicerol deshidrogenasa, familia cercanamente relacionada, de la cual se ha determinado su estructura y han sido depositadas en el PDB. Cuatro de esas secuencias son agrupadas en la subfamilia LPO con número cd08176; cuatro secuencias corresponden a la subfamilia MAR cd08177; dos secuencias se agrupan en la subfamilia BDH con número cd8187; por último hay una secuencia perteneciente a cada una de las subfamilias PPD con número cd08181 y AAD con número cd 08178.

El número de acceso de acuerdo al PDB se indica a la izquierda de cada secuencia, así como el nombre del organismo al cual pertenece, a su vez al lado derecho de cada secuencia se observa la subfamilia a la cual pertenece.

Se identificaron un total de 21 elementos de estructura secundaria, ocho de tales elementos corresponden a hebras beta y trece de tales elementos a alfa-hélices, resaltados en la imagen en colores amarillo y verde respectivamente. Los residuos que interaccionan con coenzima están resaltados en rojo. Mientras que los residuos involucrados en la unión al átomo de Fe se encuentra resaltados en morado; los residuos involucrados en la unión al átomo de zinc en las secuencias correspondientes a las glicerol deshidrogenasa se resaltan en color gris. Por último, la secuencia correspondiente a la ADH-Fe de humano se encuentra resaltada en color gris y azul correspondiente al dominio N-terminal y C-terminal respectivamente. Tal alineamiento estructural fue realizado usando la herramienta VAST de uso libre en el servidor del NCBI.

6.1 La familia de ADH-Fe o de tipo III comprende varias subfamilias de proteínas con una variedad de actividades catalíticas.

A la fecha se han reportado una variedad de proteínas reportadas como miembros de la familia ADH-Fe que exhiben una diversidad de actividades catalíticas. Además de los reportes iniciales de proteínas activadas por hierro con actividad de etanol deshidrogenasa en *Zymomonas mobilis* y *Saccharomyces cerevisiae;* se han reportado varias enzimas con otras actividades como son: metanol deshidrogenasa, lactaldehído: propanodiol oxidoreductasa (lactaldehído reductasa), propanol deshidrogenasa butanol deshidrogenasa, L-1,3-propanodiol

deshidrogenasa, maleíl acetato reductasa, L-treonina deshidrogenasa e hidroxiácido-oxoácido transhidrogenasa (Wills *et al*., 1981; Williamson *et al*.,1987; Conway *et al*., 1989; Moon *et al*., 2011; Arfman *et al*., 1997; de Vries *et al*., 1992; Vonck *et al*., 1991; Montella *et al*., 2005; Hiu *et al*., 1987; Youngleso *et al*., 1989; Daubaras *et al*., 1996; Seibert *et al*., 2004; Lyon *et al*., 2009).

De acuerdo con la base de datos de dominios conservados del NCBI, las secuencias pertenecientes a la familia de las ADH-Fe (cd08551) se encuentran distribuidas en al menos 19 distintas subfamilias resumidas en el cuadro 1 (sección de resultados). Para explorar la relación entre las distintas secuencias pertenecientes a esta familia realizamos un alineamiento con todas aquellas secuencias recuperadas de la base de datos de dominios conservados de NCBI, incorporando únicamente aquellas secuencias no redundantes cuya longitud era mayor a 200 residuos aminoácidos, tomando en cuenta que dichas secuencias tienen en promedio 470 aa en longitud. Bajo los criterios antes citados, se obtuvieron 538 secuencias, las cuales fueron usadas para realizar un análisis filogenético. Dicho análisis filogenético se muestra en la figura 3 (sección de resultados), donde se observa un árbol construido por uno de los métodos basados en caracteres más robusto, el método de máxima verosimilitud, donde se incluyen cada una de las 19 subfamilias proteicas propuestas por la base de dominios conservados del NCBI, respaldado con un alto valor de confianza de bootstrap. Adicionalmente se empleó el mejor hit reciproco en BLAST como criterio adicional, corroborando así que cada una de estas subfamilias comprende un posible grupo de proteínas ortólogas (Ward *et al*., 2014).

Por otro lado, entre las distintas subfamilias agrupadas dentro de la familia ADH-Fe, solo unas pocas presentan un alto valor de bootstrap entre ellas. Se encontró una gran diversidad de funciones dentro de esta familia, por lo tanto las secuencias que integran las distintas subfamilias han divergido más allá de niveles que pueden ser detectados por métodos estándar de alineamiento de secuencias. Aunque fue indispensable afinar dicho alineamiento, solo las subfamilias cercanamente relacionadas mostraron un alto valor de bootstrap. Por ejemplo, la

subfamilia de las lactaldehído: propanodiol oxidoreductasa (LPO) con número cd08176 se encuentra relacionada a la subfamilia FeADH4 con número cd08188 con un valor de confianza de bootstrap del 81%. Además, el dominio C-terminal de la subfamilia de doble dominio alcohol acetaldehído deshidrogenasa (AAD-C) con número (cd08178) se encuentra relacionada, respaldado por un valor de confianza de bootstrap del 95%, a la familia de butanol deshidrogenasa (BDH) con número cd08179 así como con la subfamilia propanodiol deshidrogenasa (PDD) con número cd08180.

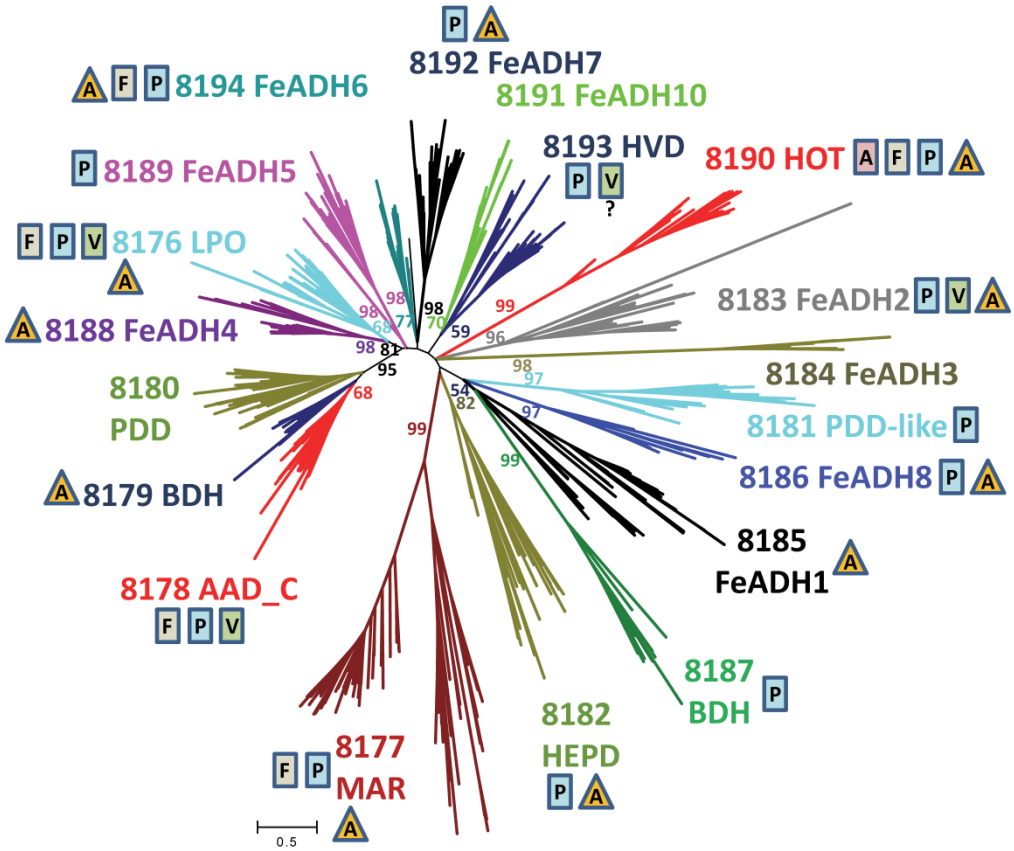


Figura 3. Análisis filogenético de 538 secuencias de ADH-Fe recuperadas de la base de datos de dominios conservados del NCBI. El árbol filogenético sin enraizar fue obtenido mediante el algoritmo de Máxima Verosimilitud tomando como modelo de sustitución el modelo de Le-Gascuel. Las ramas están coloreadas de acuerdo a la base de datos de dominios conservados a la subfamilia a la que pertenecen. Se muestra el árbol con el más alto log likelihood (-2505413,5328). Se obtuvieron árboles con topología similar con algoritmos como son: Máxima Parsimonia, Mínima Evolución y Vecino más Cercano, se empleó

una distribución gamma discreta para modelar la diferencia de la tasa evolutiva entre sitios (5 categorías (+G, parámetro = 0.8682)). El árbol fue dibujado a escala con la longitud de las ramas representando el número de sustituciones por sitio. Hubo un total de 783 posiciones en el juego de datos final. Se indica el valor de confianza de bootstrap cerca de cada rama en una prueba de 500 réplicas. Los rectángulos y triángulos adyacentes al nombre de cada subfamilia de ADH-Fe, indican la presencia de secuencias proteicas del dominio archaea (triángulos) o del dominio eukarya (rectángulos con A (animales), F (fungi), V (viridiplantae) y P otros eucariontes, en cada una de las subfamilias. Las secuencias de bacteria se encuentran presentes en todas las subfamilias de ADH-Fe.

Cuadro 1. Subfamilias que conforman la familia de alcohol deshidrogenasas dependientes de hierro.

CDD Protein Subfamily	Reported activity [properties]/ Characterized proteins [accession number]	Reported structure	Phyletic distribution			Reference
			Bacteria	Eukarya	Archaea	
LPO cd08176	**Lactaldehyde:propanediol oxidoreductase (lactaldehyde reductase)**[dimer; NAD^+-dependent activity] FucO *Escherichia coli* [P0A9S1] [crystallized either with NAD^+, Fe^{2+}, 1,2-propanediol, adenosine diphosphoribose, or Zn^{2+}]	2BL4; 2BI4; 1RRM	Yes	Yes (Euglenozoa; Heterolobosea; **Fungi**, Ascomycota; **Viridiplantae**, Chlorophyta)	Yes	[23;59]
	L-1,3-Propanediol dehydrogenase [dimer; NAD^+-dependent activity] DhaT *Klebsiella pneumonia* [Q59477][pentamer of dimers; crystallized with Fe^{2+}] DhaT *Clostridium pasteurianum* [o30454] ADH3 (*dhaT* gene) *Oenococcus oeni* [EKP90059] [dimer; crystallized with Ni^{2+}]] DhaT *Citrobacter freundii* [p45513]	 3BFJ -- 4FR2 --				 [66] [65] [64] [63]
	Methanol dehydrogenase [NAD^+-dependent activity] MDH *Bacillus methanolicus* [P31005] [pentamer of dimers; contains Zn^{2+} and Mg^{2+}]	 --				 [56-58]
	Ethanol dehydrogenase [NAD^+-dependent activity] ADH2 *Zymomonas mobilis* [F8DVL8] [crystallized with NAD^+, Fe^{2+}] ADH4 *Saccharomyces cerevisiae* [P10127][dimer][Zinc activated enzyme] ADH4 Schizosaccharomyces pombe [Q09669]	 3OX4 -- --				 [18;23;55] [21;23] [117]
	L-threonine dehydrogenase [NAD^+-dependent activity] YiaY *Escherichia coli* [P37686] [enzyme contain both Fe^{2+} and Zn^{2+}]	 --				 [72]
MAR cd08177	**Maleylacetate reductase** [NADH dependent activity] Involved in the degradation of substituted aromatic compounds through the 3-oxoadipate pathway MacA1 Rhodococcus opacus 1CP [O84992]	 --	Yes	Yes (Haptophyceae; Stramenopiles; **Fungi**, ascomycota,	Yes	 [70]

CDD Protein Subfamily	Reported activity [properties]/ Characterized proteins [accession number]	Reported structure	Phyletic distribution			Reference
			Bacteria	Eukarya	Archaea	
	TfdFI (Reut_D6463) *Cupriavidus necator* JMP134 (previously known as *Ralstonia eutropha* or *Alcaligenes eutrophus*) [P27137]	--		basidiomycota)		[68;69]
	TfdFII (Reut_D6471) *Cupriavidus necator* JMP134 [P94135]	--				[68;69]
	TcpD (Reut_A1589) *Cupriavidus necator* JMP134 [Q471H8; AAZ60955]	--				[68;69]
	HxqD (Reut_B4129) *Cupriavidus necator* JMP134 [Q46TQ1] [crystallized with NAD^+]	3JZD				[68;69]
	HqoD (Reut_B4694) *Cupriavidus necator* JMP134 [Q46S41]	--				[68;69]
	MAR (Ncgl2952 locus) *Corynebacterium glutamicum* ATCC 13032 [Q8NL91; NP_602249]	3IV7				[90]
	MAR (Ncgl1112 locus) *Corynebacterium glutamicum* ATCC 13032 [Q8NR93; NP_600385]	--				[90]
	DxnE (Swit_4891 locus) *Sphingomonas wittichii* RW1 [A5VGV4; ABQ71513]	--				[136]
	TftE *Burkholderia cepacia* AC1100 [Q45072]	--				[67;137]
	LinF *Sphingobium japonicum* UT26 (formerly *Sphingomonas paucimobilis* UT26) [Q5W9E3; BAD66863]	--				[138]
	CcaD *Pseudomonas reinekei* MT1 [C6YXH0; ABO61029]	--				[139]
	GraC *Rhizobium* sp. MTP-10005 [A1IIX4; BAF44524] [homodimer]	3W5S				[140]
	MacA (Atu2528 locus) *Agrobacterium fabrum* str. C58 [NP_355474] [crystallized with NAD^+]	3HL0				Unpublished
	FUM7 *Fusarium verticillioides* Is a gen associated with fumonisin biosynthesis	--				[131]
AAD-C cd08178	**C-terminal alcohol dehydrogenase domain of the acetaldehyde dehydrogenase-alcohol dehydrogenase bifunctional two-domain protein** [NAD(H) dependent activity]		Yes	Yes (Amoebozoa; Alveolata; Diplomonadida; Cryptophyta; **Viridiplantae**, Chlorophyta; **Fungi**, ascomycota, neocallimastigomycota	No	
	ADHE Geobacillus thermoglucosidasius NCIMB 11955 [WP_013877698; C-terminal domain, 435-867 aa] [crystallized with Zn^{2+}]	3ZDR				[78;98]
	ADH2 *Entamoeba histolytica* [Q24803]	--				[77-79]
	ADHE *Escherichia coli* K-12 [P0A9Q7]	--				[80;81]

CDD Protein Subfamily	Reported activity [properties]/ Characterized proteins [accession number]	Reported structure	Phyletic distribution			Reference
			Bacteria	Eukarya	Archaea	
	Also possess activity as pyruvate-formatelyase deactivase			)		
NADPH-BDH cd08179	**NADPH-dependent butanol dehydrogenase** (*Clostridium saccharobutylicum* and *C. beijerinckii* use both ethanol and butanol as substrates) AdhA *Clostridium beijerinckii* NRRL B592 [AAM18705] ADH1 *Clostridium saccharobutylicum*, formerly *C. acetobutylicum* [P13604]	 -- --	Yes	No	Yes	 [61] [62;141]
PDD cd08180	**1,3-propanediol dehydrogenase** [NAD^+-dependent activity] ***PduQ*** *Salmonella typhimurium* LT2 [Q9XDN0; NP_460997]	 --	Yes	No	No	 [60]
PDD-like cd08181	**Putative 1,3-propanediol dehydrogenase-like** [The enzyme bound $NADP^+$] TM0920 gene of *Thermotoga maritima* [Q9X022; WP_004080642] [crystallized with $NADP^+$, Fe^{3+}, or Zn^{2+}]	 1O2D; 1VHD	Yes	Yes (Diplomonadida)	No	 [83]
HEPD cd08182	**Hydroxyethylphosphoate dehydrogenase or phosphonoacetaldehyde reductase.** Encoding gene is located inside an operon involved in the biosynthesis of phosphinothricin tripeptide (PTT), an antibiotic used as herbicide. phpC *Streptomyces viridochromogenes* DSM 40736 [AAU00078] fomC Streptomyces fradiae [ACG70833]	 -- --	Yes	Yes[5] (Stramenopiles)	Yes	 [142;143] [144;145]
BDH cd08187	**Butanol dehydrogenase / aldehyde reductase** $NADP^+$-dependent activity with preference for alcohols longer than C3 YqhD Escherichia coli [Q46856] [crystallized with $NADP^+$, Zn^{2+}] BDH (TM0820 locus) *Thermotoga maritima* MSB8 [NP_228629; Q9WZS7] [crystallized with $NADP^+$, Fe^{3+}]	 1OJ7; 4QGS 1VLJ	Yes	Yes (Stramenopiles; Amoebozoa; Parabasalidea)	No	 [84] Unpublished
HOT cd08190	**Hydroxyacid-oxoacid transhydrogenase** ADHFe1 Homo sapiens [Q8IWW8]	 --	Yes	Yes (Ichthyosporea; Apusozoa; Stramenopiles;	Yes	 [28;73;102;105;146]

CDD Protein Subfamily	Reported activity [properties]/ Characterized proteins [accession number]	Reported structure	Phyletic distribution			Reference
			Bacteria	Eukarya	Archaea	
	ADHFe1 Rattus norvegicus [Q4QQW3]	--		Amoebozoa; Rhizaria; **Fungi**; **Metazoa**)		
HVD cd08193	**5-hydroxyvalerate dehydrogenase** CpnD *Comamonas* sp. NCIMB 9872 [BAC22648]		Yes	Yes[4] (Haptophyceae ; **Viridiplantae**, tracheophyta)	No	[147]
FeADH1 cd08185	None characterized		Yes	No	Yes	
FeADH2 cd08183	None characterized		Yes	Yes (Alveolata; Stramenopiles; Rhodophyta; **Viridiplantae,** Chlorophyta)	Yes[1]	
FeADH3 cd08184	**3-deoxy-alpha-D-manno-octulosonate 8-oxidase** Catalyzes the first step of the biosynthesis of Kdo8N (8-amino-3,8-dideoxy-D-manno-octulosonate), found in lipopolysaccharides of members of the Shewanella genus KdnB *Shewanella oneidensis* [Q8EEB0]		Yes	No	No	[148]
FeADH4 cd08188	None characterized		Yes	No	Yes	
FeADH5 cd08189	None characterized		Yes	Yes (Euglenozoa)	No	
FeADH6 cd08194	None characterized		Yes	Yes (Alveolata; Haptophyceae; Rhizaria; Ichthyosporea; **Fungi**, chytridiomycota)	Yes	
FeADH8 cd08186	**$NADP^+$-dependent ADH**, oxidizes a series of primary aliphatic and aromatic alcohols, but shows a higher affinity for aldehyde substrates. *Thermococcus paralvinellae* (formerly T. sp. ES1) [ACK56133]	--	Yes	Yes Diplomonadida	Yes	[85-88]

CDD Protein Subfamily	Reported activity [properties]/ Characterized proteins [accession number]	Reported structure	Phyletic distribution			Reference
			Bacteria	Eukarya	Archaea	
	Thermococcus sp. AN1 [AAB63011] *Thermococcus hydrothermalis* [CAA74334]	-- --				
FeADH7 cd08192	**Iron-containing alcohol dehydrogenases** probably involved in the linear alkylbenzenesulfonate (LAS) degradation pathway [in *Parvibaculum lavamentivorans* the expression of the gene encoding this enzyme is induced during growth with LAS] [2]. *Parvibaculum lavamentivorans* [ABS64400]		Yes	Yes (Haptophyceae; Stramenopiles)	Yes[3]	
HHD FeADH10 cd08191	**6-hydroxyhexanoate dehydrogenase** ChnD1 Brevibacterium sp. HCU [AAK73161]		Yes	No	No	[149]

1 Found in Lokiarchaeum sp. GC14_75.

2 Personal communication from Dr. David Schleheck, University of Konstanz.

3 Found in Thaumarchaeota archeon SCGC AB-539-E09.

4 Found in Posidonia oceanica, a Mediterranean seagrass.

5 Found in Nannochloropsis gaditana, an oleaginous microalgae.

6.2 Distribución Filética

Los resultados obtenidos muestran que las secuencias agrupadas en la familia ADH-Fe pertenecen a organismos de los tres dominios de la vida; arqueas, bacterias y eucariontes, como se observa en el cuadro 1 (sección de resultados). En eucariontes los miembros de la familia ADH-Fe presentan una amplia distribución, por lo que pueden ser encontrados en los distintos reinos de la vida como son Plantae, Fungi y Animalia, además de estar bien representados en el grupo de Protoctistas. En la figura 4 (sección de resultados) se muestra un árbol filogenético que agrupa todas las secuencias proteicas pertenecientes a las distintas subfamilias de ADH-Fe en eucariontes, obteniendo un total de 868 secuencias, de las cuales 656 pertenecen a la subfamilia HOT con número cd08190, por lo que dicha subfamilia es la que mejor representada se encuentra en los eucariontes con un 75% del total. Una mayor distinción puede hacerse todavía, ya que las 306 secuencias proteicas recuperadas pertenecientes a los animales se encuentran agrupadas en la subfamilia HOT, mientras que 306 secuencias de 334 en hongos, es decir el 80% se agrupan también dentro de dicha subfamilia. Otros eucariontes con proteínas englobadas dentro de esta subfamilia son organismos del grupo amoebozoa como *Acanthamoeba castellani, Polysphondylium pallidum, Acytostelium subglobosum, Dictyostellium discoideum, D. lacteum,* y *D. purpureum;* del grupo estramenopiles como *Phaeodactylum tricornutum, Thalassiosira oceanica, Aphanomyces astaci, A. invadens, Saprolegnia parasitica* y *S. diclina;* del grupo icthyosporea como *Capsaspora owczarzaki;* del grupo Apuzoa como *Thecamonas trahens;* y del grupo Rhizaria como el foraminífero *Reticulomyxa filose.* Cabe mencionar que los organismos del reino Plantae no estuvieron representados en la subfamilia HOT.

Por otro lado, las secuencias en hongos de las ADH-Fe se encuentran clasificadas en cinco subfamilias, el 80% de tales secuencias pertenecen a la subfamilia HOT, aparentemente distribuidas en todos los grupos del reino Fungi con excepción de los saccharomycetes; que incluye levaduras de importancia comercial como lo son *Saccharomyces cerevisiae* y *Kluyveromyces lactis;* y schizosaccharomycetes

como *Schizosaccharomyces pombe*; otra subfamilia en la cual se agrupan las secuencias de Fungi corresponde a la subfamilia MAR con número cd08177, la cual incluye casi el 17% de las secuencias pertenecientes a este reino de la vida, las cuales pertenecían casi en su totalidad a ascomicetos y basidiomicetos. Además, todas las secuencias de ADH-Fe reportadas de saccharomycetes y schizosaccharomycetes se agrupan en la subfamilia LPO con número cd08176, estando involucradas en el metabolismo del etanol. Casi para finalizar, tenemos tres secuencias de Fungi que pertenecen a la subfamilia AAD-C como son; *Togninia minima* del grupo ascomycota y *Neocallimastix frontalis* y *Piromyces sp.* E2 del grupo neocallimastigomycota. Cabe mencionar que en esta última subfamilia presentan un dominio correspondiente a las ADHE, el cual se ha propuesto resultó de la transferencia horizontal de genes por parte de diferentes bacterias (Andersson *et al*., 2006). Por último, se encuentra el curioso caso de la ADH-Fe perteneciente a la subfamilia cd08194 encontrada en *Gonapodya prolifera* JEL478 perteneciente al grupo de los quitridiomicetos, dicha secuencia es difícil de explicar por transferencia horizontal de genes dado que el gen codificante para dicha proteína presenta 9 exones. Dificultando las especulaciones sobre el origen de esa única secuencia en el reino Fungi.

Las ADH-Fe se encuentran ausentes en plantas superiores, encontrándose únicamente en algas verdes, curiosamente en el grupo de las clorofitas, grupo donde se encuentra clasificado el último ancestro común entre algas y plantas terrestres, por lo que la ausencia de tal gen en embriofitas sugiere que dicho gen fue perdido en el último ancestro común de dicho grupo taxonómico. Contrastando con esto último, los diferentes grupos de clorofitas *sensu lato* poseen ADH-Fe pertenecientes a distintas subfamilias proteicas, como puede analizarse en el cuadro 2 (sección de resultados). De esta forma, los taxa de la clase Chlorophyceae poseen únicamente ADH-Fe pertenecientes a la subfamilia de proteínas bidominio alcohol y acetaldehído deshidrogenasa (AAD-C). Mientras que algas de la clase Trebouxiophyceae poseen ADH-Fe que pertenecen a esta misma subfamilia AAD-C y a la subfamilia de lactaldehído: propanodiol oxidoreductasa (LPO), por último, las algas de la clase Prasinophyceae poseen

ADH-Fe pertenecientes a subfamilias como LPO (cd08176) y una subfamilia de proteínas no caracterizada que presenta el número cd08183 en la base de datos de dominios conservados en el NCBI.

Para concluir el apartado correspondiente a la distribución de tales secuencias en plantas, debe mencionarse que en 2012 una ADH-Fe perteneciente a la subfamilia de 5-hidroxivalerato deshidrogenasa (HVD) con número cd08193, ha sido reportada en una embriofita llamada *Posidonia oceanica*, también conocida como pasto marino del mediterráneo. Tal proteína fue secuenciada de un mRNA aislado que cambio su expresión en respuesta a tratamiento con cadmio; mostrando la más alta identidad, del 65%, con la ADH-Fe de *Rhizobium leguminosarum*, una α-proteobacteria. No tenemos certeza si esta secuencia de ADH-Fe encontrada en *Posidonia oceanica* pudo haber surgido a través de una transferencia horizontal de genes de una bacteria hacia la planta, en un proceso posible pero poco probable, aunque se encuentran ejemplos similares documentados en la literatura. Por otro lado, podría tratarse únicamente de contaminación bacteriana a la hora del procedimiento de aislamiento total de RNA de hojas y yemas apicales, esto último tendría como fundamento que bacterias del genero *Rhizobium* frecuentemente se encuentran formando asociación simbiótica con la raíz de distintas plantas. Debido a que en ninguna otra embriofita ha sido reportada a la fecha, la presencia de otra secuencia de ADH-Fe y no existiendo más información citada en la literatura que la aquí mostrada, la presencia de tal secuencia en dicha embriofita debe ser cuestionada (Greco *et al*., 2012).

Es interesante que Fungi y las Clorofitas exhiben una distribución irregular de las ADH-Fe y aún más de llamar la atención es el hecho de que las ADH-Fe pertenecientes a las diferentes subfamilias proteicas no son funcionalmente equivalentes, ya que participan en distintas funciones metabólicas.

Son de esperarse variaciones entre homólogos remotos entre las distintas subfamilias. Ya que mientras la especificidad de sustrato puede ser diversa dentro de una misma familia, el mecanismo de reacción se mantiene más a menudo dentro de ésta.

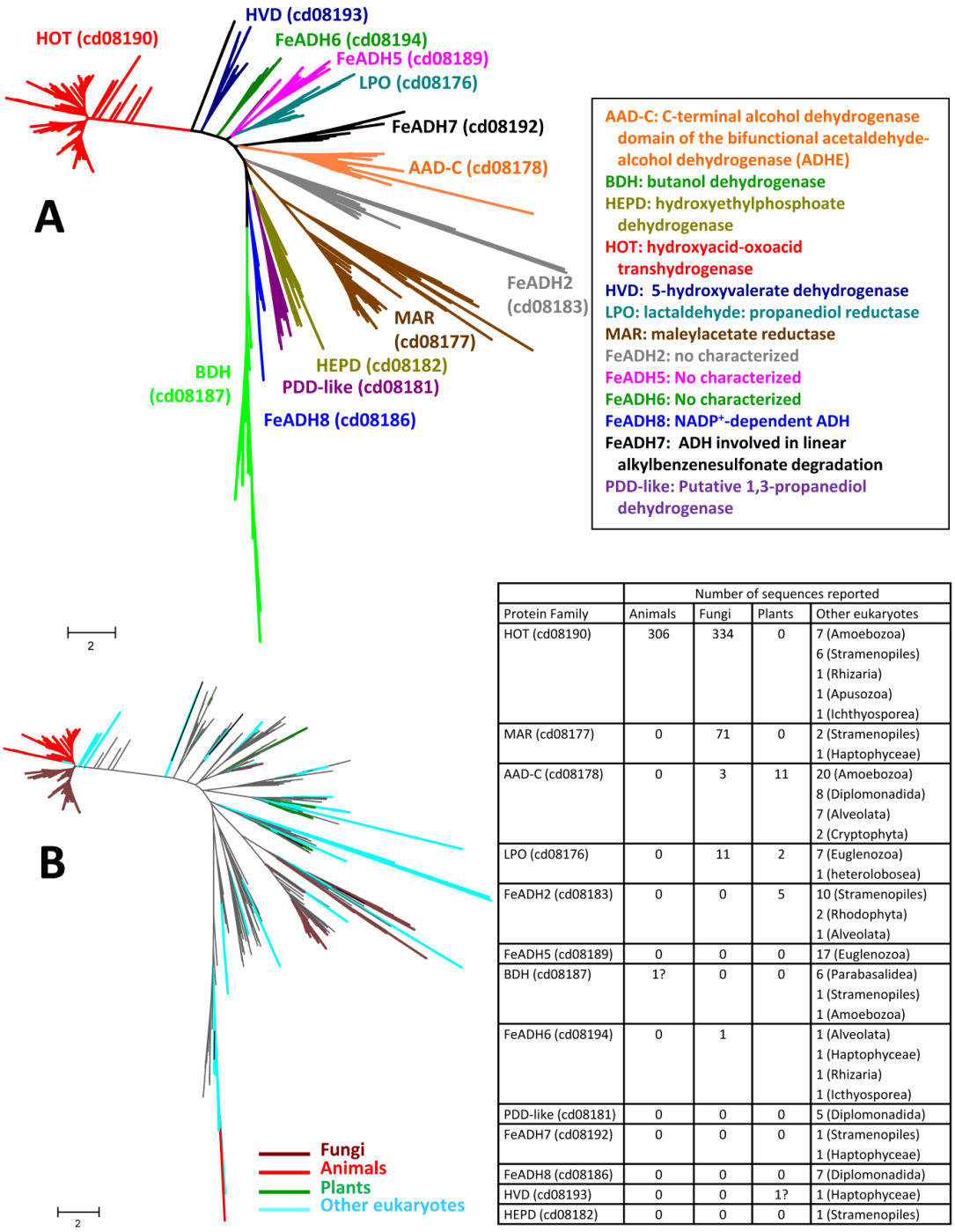
No se trata de un caso aislado, existen numerosos casos donde las proteínas clasificadas en una familia han divergido a niveles que son difíciles de detectar por métodos de alineamiento de secuencias estándar. Necesitándose en aquellos casos el uso de alineamientos estructurales, como el empleado en el presente estudio.

Un ejemplo lo tenemos aunque a nivel de superfamilia en las haloácido dehalogenasas, enzimas que pueden procesar una vasta variedad de sustratos, pero que siempre actúan vía la formación de un intermediario enzima-sustrato por medio de un residuo de ácido aspártico conservado, el cual se encarga de facilitar la escisión de enlaces C-Cl, P-C o P-O. En este caso particular, cada una de las 20 familias de las haloácido dehalogenasas, cataliza una reacción única.

La presencia de un paso funcional común dentro de familias o superfamilias con una variedad de funciones, apunta hacia un mantenimiento del mecanismo catalítico de estas enzimas en el curso de la diversificación. Tal situación da un indicio de un escenario evolutivo en el cual las enzimas adquirían nuevas funciones vía duplicación y reclutamiento, mediante el mantenimiento aunque de forma parcial de los mecanismos de reacción en vez de la especificidad de sustrato, dando como resultado familias donde la diversidad de función es alta.

Otro ejemplo lo encontramos en la familia de ADH-Zn donde de acuerdo a los análisis filogenéticos, las secuencias de dicha familia derivan de las enzimas de la clase III (ADH-3) mediante procesos de duplicación génica. Las enzimas de las clases restantes fueron reclutadas para realizar una diversidad de funciones procesando diversos metabolitos endógenos (Sanghani *et al*., 2002; Godoy 2006 *et al*., 2006; Jörnvall *et al*., 1970).

Por último tenemos el caso de las ADH de cadena corta, las cuales de acuerdo a análisis filogenéticos realizados en 2011, se propone divergieron de las prostaglandinas deshidrogenasas (PGDH). Después de un proceso de duplicación de los genes de estas últimas.



	Number of sequences reported			
Protein Family	Animals	Fungi	Plants	Other eukaryotes
HOT (cd08190)	306	334	0	7 (Amoebozoa) 6 (Stramenopiles) 1 (Rhizaria) 1 (Apusozoa) 1 (Ichthyosporea)
MAR (cd08177)	0	71	0	2 (Stramenopiles) 1 (Haptophyceae)
AAD-C (cd08178)	0	3	11	20 (Amoebozoa) 8 (Diplomonadida) 7 (Alveolata) 2 (Cryptophyta)
LPO (cd08176)	0	11	2	7 (Euglenozoa) 1 (heterolobosea)
FeADH2 (cd08183)	0	0	5	10 (Stramenopiles) 2 (Rhodophyta) 1 (Alveolata)
FeADH5 (cd08189)	0	0	0	17 (Euglenozoa)
BDH (cd08187)	1?	0	0	6 (Parabasalidea) 1 (Stramenopiles) 1 (Amoebozoa)
FeADH6 (cd08194)	0	1		1 (Alveolata) 1 (Haptophyceae) 1 (Rhizaria) 1 (Icthyosporea)
PDD-like (cd08181)	0	0	0	5 (Diplomonadida)
FeADH7 (cd08192)	0	0	0	1 (Stramenopiles) 1 (Haptophyceae)
FeADH8 (cd08186)	0	0	0	7 (Diplomonadida)
HVD (cd08193)	0	0	1?	1 (Haptophyceae)
HEPD (cd08182)	0	0	0	1 (Stramenopiles)

Figura 4. Análisis filogenético de las 867 secuencias de ADH-Fe de eucariontes más 352 secuencias no redundantes recuperadas de la base de datos de dominios conservados del NCBI. El análisis filogenético fue realizado con el algoritmo de Máxima Verosimilitud, tomando como modelo de sustitución el modelo de Le-Gascuel. Se muestra el árbol con el log más alto (-3414819.0869). El o los arboles iniciales para la búsqueda heurística fueron obtenidos aplicando

automáticamente los algoritmos Neighbor-Joining y BioNJ a una matriz de comparación de distancias estimadas usando el modelo JTT y luego seleccionando la topología con mayor valor de verosimilitud de log. Una distribución gamma discreta fue usada para modelar las diferencias en las tasas evolutivas entre sitios (5 categorías (+G, parámetro =283 0.4901)). El árbol se dibujó a escala, con la longitud de la rama representando el número de sustituciones por sitio. El análisis incorporó un total de 1219 secuencias de aminoácidos. Hubo un total de 996 posiciones en el juego de datos final.

Cuadro 2. Número de secuencias de ADH-Fe de plantas encontradas en las distintas subfamilias.

Organism	Number of genes			
	cd08178 AAD-C	cd08176 LPO	cd08183 FeADH2	cd08193 HVD
Chlorophyta				
Class Chlorophyceae				
Chlamydomonas reinhardtii (CC-503 cw92 mt+)	3	0	0	0
Polytomella sp. Pringsheim 198.80	1	0	0	0
Volvox carteri f. Nagariensis	2	0	0	0
Gonium pectoral	2	0	0	0
Monoraphidium neglectum SAG 48.87	1			
Class Trebouxiophyceae				
Chlorella variabilis NC64A	2	0	0	0
Auxenochlorella protothecoides 0710 (*Chlorella protothecoides*)	0	1	0	0
Class Prasinophyceae				
Micromonas pusilla CCMP1545	0	1	1	0
Micromonas sp. RCC299 (*Micromonas commoda*)	0	0	1	0
Ostreococcus tauri	0	0	1	0
Ostreococcus lucimarinus CCE9901	0	0	1	0
Bathycoccus prasinos	0	0	1	0
Streptophyta, Tracheophyta (monocot)				
Posidonia oceánica (Mediterranean seagrass)	0	0	0	1?

6.3 Las ADH-Fe comparten el mismo plegamiento

Se ha resuelto la estructura tridimensional de 12 secuencias de ADH-Fe, estas estructuras pueden ser agrupadas en cinco de las diferentes subfamilias pertenecientes a la familia de las ADH-Fe (cd08551). Al analizar las estructuras resueltas, podemos observar que todas están formadas por dos dominios, los cuales se encuentran separados por un surco profundo. El dominio ubicado en el extremo N-terminal contiene estructuras secundarias correspondientes a α hélices y a hebras β, formando un motivo de unión a nucleótido, conocido como plegamiento tipo Rossmann, donde se une la coenzima. En cuanto al dominio C-terminal, éste se encuentra compuesto exclusivamente de estructuras secundarias correspondientes a 9 α hélices donde se ubica el sitio de unión a hierro.

Como se observa en la figura 5 (sección de resultados), dicho plegamiento se encuentra conservado en todos los miembros de la familia de proteínas ADH-Fe y también en familias relacionadas como son la familia glicerol deshidrogenasa (GDH) (cd08550), la familia glicerol-1-fosfato deshidrogenasa (G1PDH) (cd08549), y la familia dehidroquinato sintasa (DHQ) (cd08169), todas ellas pertenecientes a la superfamilia DHQ-FeADH (cd07766). Sin embargo, vale la pena mencionar que en esta última familia (DHQ), el dominio C-terminal comprende además de α-hélices entre dos y cuatro hebras β. La identidad de secuencia entre proteínas que pertenecen a diferentes subfamilias de ADH-Fe es casi del 20% (entre un 30 y 40% de similitud). En contraste, la identidad de secuencia entre, proteínas que pertenecen a diferentes familias proteicas dentro de la superfamilia DHQ-FeADH (cd07766) es casi del 10% y una similitud del 20%. Por lo tanto, aunque las secuencias proteicas de las distintas subfamilias son altamente divergentes dentro de la superfamilia DHQ-FeADH, en la figura 5 (sección de resultados) se puede observar que todas ellas comparten un plegamiento y similitud de dominios.

En la figura 2 (sección de resultados) se muestra un alineamiento múltiple de secuencias de ADH-Fe con estructura tridimensional resuelta. Se puede observar que las 21 estructuras secundarias que comprende el plegamiento de las ADH-Fe

se encuentran muy bien conservadas en todas las ADH-Fe reportadas, distribuidas como 8 hebras β y 13 α-hélices. El dominio N-terminal comprende los residuos 1-229, tomando como base la numeración de la ADH-Fe humana, concentrando 9 de las 13 α-hélices. Por lo tanto dicho dominio se encuentra involucrado en la unión a coenzima NAD(P)H; mientras que el dominio C-terminal posee los aminoácidos conservados que tienen que ver con el enlace de coordinación con el ion metálico.

Con motivos de comparación, cuatro estructuras de la familia relacionada glicerol fosfato deshidrogenasa, cd08550, fueron incluidas en el análisis, mostrando una distribución de estructuras secundarias similar, con algunas ligeras diferencias como son; el asa localizada entre el α hélice número 4 y la hebra β número 5 son más cortas en las glicerol deshidrogenasa, mientras que la α hélice número 6 se encuentra desplazada 8 residuos de su posición y las α hélices 7 y 8 se encuentran unidas formando una sola hélice. Todas estas diferencias, incluyendo los datos de la figura 1 (sección de resultados), apoyan la idea de que las glicerol deshidrogenasa deben ser consideradas como una familia proteica separada de la familia ADH-Fe *bona fide*.

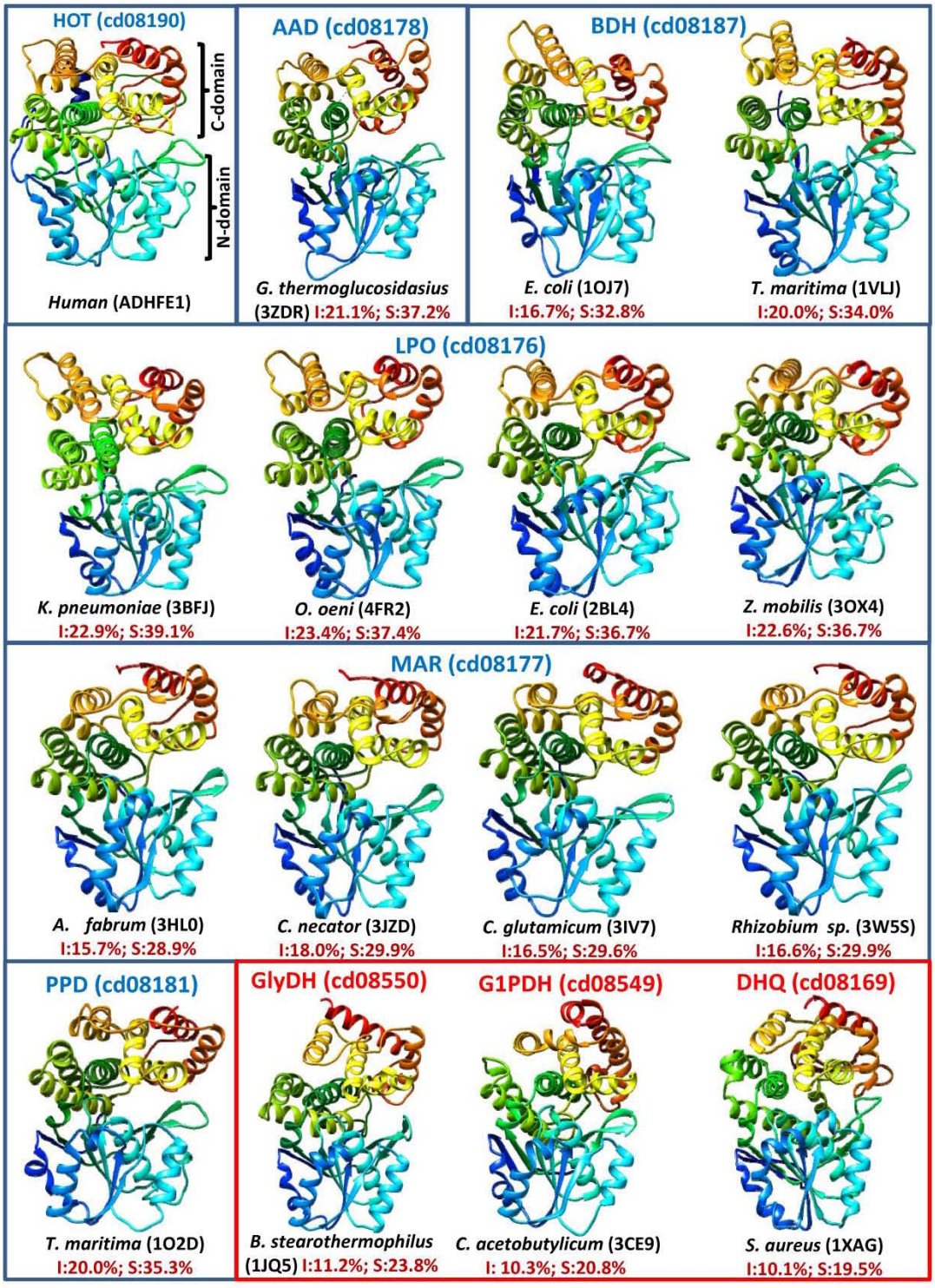


Figura 5. Comparación de distintas ADH-Fe de las cuales se ha reportado su estructura tridimensional en el PDB. Todas las estructuras que han sido resueltas se agrupan en 5 subfamilias de las ADH-Fe (indicando en azul la subfamilia a la que pertenecen, en la parte superior de cada estructura). En la parte inferior de cada recuadro el nombre científico del organismo en color negro, al cual pertenece

cada proteína, además del número de acceso en el PDB indicado en paréntesis del mismo color. Las últimas tres estructuras, rectángulo rojo, corresponden a proteínas resueltas agrupadas en familias homólogas cercanas dentro de la superfamilia DHQ_ FeADH (cd07766). La primera estructura, parte superior izquierda, corresponde a una predicción hecha con el servidor I-TASSER, método de predicción de la estructura terciaria por enhebrado, de la ADH-Fe de humano (número de ascensión NP_653251) la cual, como se ha discutido a lo largo de la sección de resultados, pertenece a la subfamilia HOT, junto con el resto de las secuencias de animales. Los números en color vino debajo de cada una de las estructuras muestran los valores de identidad (I) y (S) similitud entre las distintas secuencias tomando como base de cada comparación la ADH-Fe de humano. Los elementos de estructura secundaria fueron coloreados de acuerdo a la opción colorear “rainbow” del software UCSF Chimera versión 1.9. Dicha opción emplea los colores del arcoíris para los elementos de estructura secundaria que encuentra, tomando como punto de partida el extremo N-terminal. Todas las estructuras mostradas en recuadro azul, a excepción de la predicción hecha por I-TASSER, fueron empleadas para realizar nuestro alineamiento estructural, el cual posteriormente sirvió como base de nuestro alineamiento múltiple de secuencias y posteriormente del análisis filogenético. Cuatro de esas estructuras se agrupan en la subfamilia LPO con número cd08176; cuatro estructuras corresponden a la subfamilia MAR cd08177; dos estructuras a la subfamilia BDH con número cd8187; por último una estructura perteneciente a cada una de las subfamilias PPD con número cd08181 y AAD con número cd 08178.

6.4 Sitio de unión a coenzima

Como se describió en la sección previa, el dominio N-terminal alberga un plegamiento tipo Rossmann que contiene el sitio de unión a coenzima. Los residuos de unión a coenzima se encuentran conservados en todas las subfamilias de ADH-Fe. También se encontró un motivo GGGS que corresponde a los residuos 138 a 141 tomando como base la secuencia humana, el cual está bien conservado en todas las subfamilias, como se ejemplifica en la figura 2 (sección de resultados). Este motivo interactúa con el grupo pirofosfato del NAD(P), formando un asa que conecta la hebra β 4 y el α hélice 4.

Existe poca evidencia experimental sobre la preferencia de coenzima por parte de las ADHFe, pero existen datos que sugieren que la especificidad está determinada por la naturaleza de un aminoácido ubicado en la posición 81, tomando como base

la ADHFe humana. Adicionalmente, ocho distintas enzimas han sido cristalizadas a la fecha, unidas a coenzima; obteniendo que aquellas enzimas que usan NAD^+ poseen como aminoácido al aspartato o a la treonina en la posición antes mencionada. Por otro lado, las enzimas que unen $NADP^+$ poseen una glicina en dicha posición. En la figura número 6 (sección de resultados) se muestra la conservación de los residuos en la posición 81 a través de un análisis de logos.

La posible explicación a lo descrito en el párrafo previo es que las enzimas que presentan aspartato en la posición 81 tienen una preferencia por el NAD^+ como coenzima, debido ya sea a que la cadena lateral de dicho aminoácido repele electrostáticamente el grupo 2'-fosfato del $NADP^+$ o bien a impedimentos de tipo estérico entre estos dos mismos elementos. Además, el grupo carboxilo del aspartato interactúa directamente con el grupo hidroxilo del carbono 2 del azúcar ribosa de la adenina. Ejemplos son: la enzima FucO de *E. coli*, DhaT *de Klebsiella pneumoniae*, ADH2 de *Zymomonas mobilis*, MDH de *Bacillus methanolicus*, ADH4 de *Saccharomyces cerevisiae*, ADH2 de *Entamoeba histolytica* y ADHE de *E. coli*.

Por otro lado, la cadena lateral corta de la glicina, cuando ésta se ubica en la posición 81, se encuentra alejada de la ribosa, dejando espacio para la unión del grupo 2'-fosfato del $NADP^+$. Por lo tanto, enzimas que presentan un aminoácido con una cadena lateral corta en dicha posición, pueden unir ya sea NAD^+ o $NADP^+$ presentando algunas veces mayor afinidad por este último. Ejemplos de enzimas con glicina en la posición 81 que presentan unida $NADP^+$ son; 1,3-propanodiol deshidrogenasa perteneciente a *Thermotoga marítima*, YqhD de *E. coli*, butanol deshidrogenasa de *Thermotoga marítima* (PDB: 1VLJ), y ADHFe de *Thermococcus hydrothermalis*, *T. paralvinellae* y *T. sp.* AN1.

La serina y treonina son dos aminoácidos de cadena lateral corta que han sido asociados en otras enzimas dependientes de $NADP^+$, como residuos que pueden unirse al grupo 2'-fosfato del $NADP^+$. Sin embargo, HxqD de *Cupriavidus necator* JMP134, y MacA de *Agrobacterium fabrum*, son dos enzimas pertenecientes a la subfamilia de maleíl acetato reductasa (cd08177) las cuales fueron cristalizadas con NAD^+ como coenzima (PDB: 3JZD and 3HL0), donde ambas poseen treonina

en la posición 81. Adicionalmente, la maleíl acetato reductasa (Ncgl1112) de *Corynebacterium glutamicum* puede usar ambas coenzimas NAD^+ y $NADP^+$, a pesar de que esta enzima tiene una glicina en la posición 81. A pesar de que el residuo localizado en la posición 81 es el principal determinante sobre la preferencia de una coenzima, residuos adicionales deben estar implicados. Dicha conclusión ha sido también obtenida al analizar otras enzimas que emplean ambas coenzimas, como por ejemplo las aldehído deshidrogenasas.

Por otro lado, González Segura y cols. (González *et al*., 2015) analizaron la preferencia de las distintas familias de aldehído deshidrogenasa (ALDH) sobre una coenzima específica, encontrando la peculiaridad de que la preferencia de coenzima es una característica variable dentro de muchas familias de aldehído deshidrogenasas. En todas ellas dependía de un solo residuo el cual al parecer no presenta ningún otro papel ya sea funcional o estructural, el cual puede por lo tanto, ser fácilmente cambiado a través de la evolución y seleccionado en respuesta a necesidades fisiológicas. Podemos extrapolar dicha conclusión a las distintas subfamilias de ADHFe, dado que de igual forma, el aminoácido 81 no se encuentra conservado en algunas subfamilias como son: la subfamilia MAR, subfamilia HEPD, la subfamilia FeADH2 y la subfamilia FeADH8. Es probable que en tales subfamilias la preferencia de coenzima sea una característica variable como pasa en las familias de ALDH.

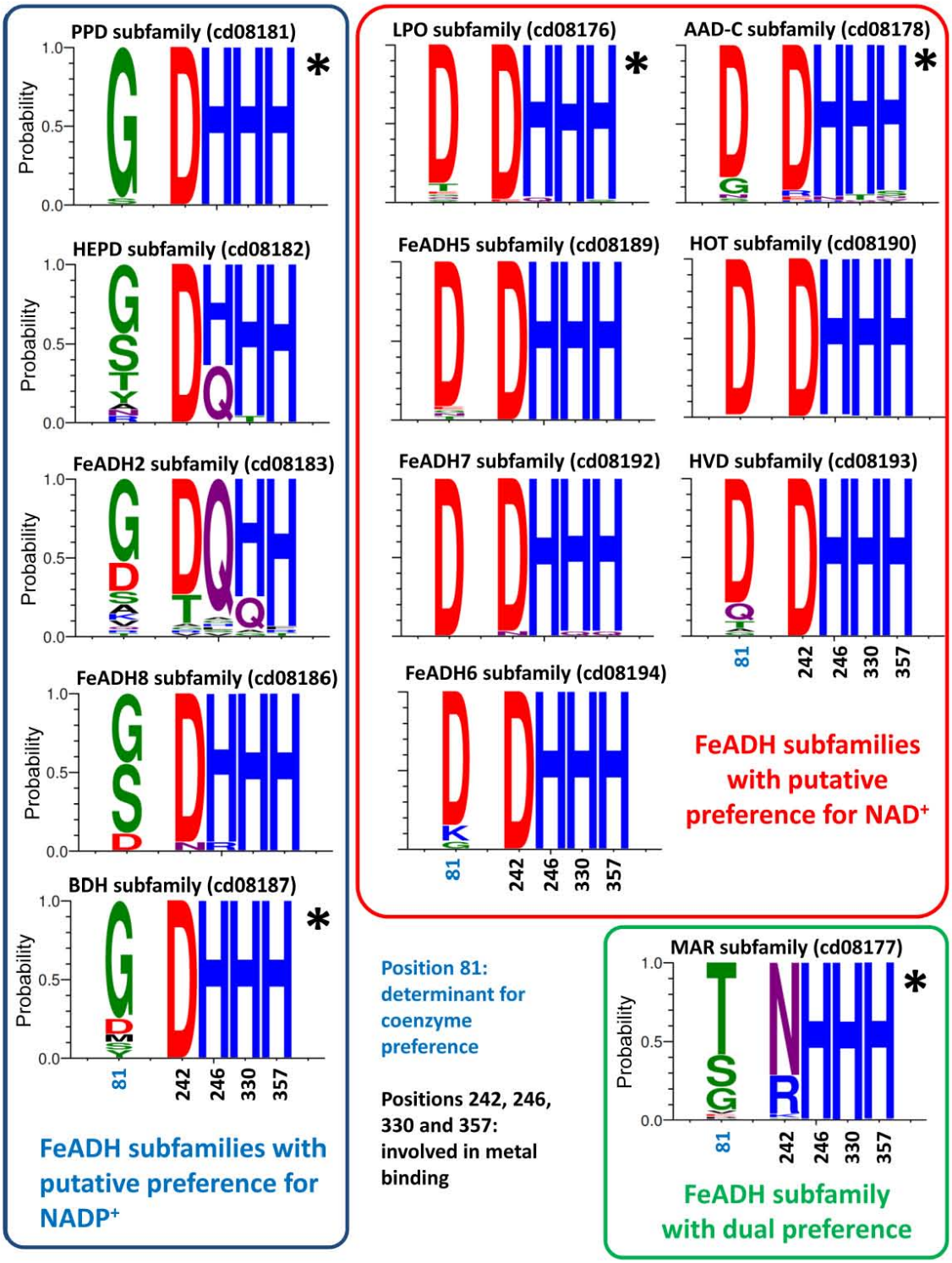


Figura 6. Análisis de LOGOS para ciertas regiones de importancia de secuencias pertenecientes a las distintas subfamilias de las ADH-Fe. Los residuos en la posición 81 son claves en la preferencia de la coenzima; mientras que residuos en las posiciones 242, 246, 330 y 357 están involucrados en la unión al ion metal. Los aminoácidos en el esquema están coloreados de acuerdo a sus

propiedades químicas: polares (G, S, T, Y, C), verde; neutrales (Q, 426 N), morado; básicos (K, R, H), azul; ácidos (D, E), rojo; e hidrofóbicos (A, V, L, 427 I, P, M, W, F), negro. Las subfamilias de ADH-Fe cuyos miembros probablemente usen NADP están englobadas en cuadro azul, aquellos que usan NAD en cuadro rojo y aquellas que usan ambas coenzimas en cuadro verde. Aquellas familias para las que existe evidencia experimental se marcan con asterisco. Los diagramas de LOGOS fueron hechos usando WebLogo 3 (http://weblogo.threeplusone.com). ♣ se tomó como base de la numeración la secuencia de ADH-Fe humana.

6.5 Sitio de unión a metal

La mayoría de las subfamilias de ADHFe, presentan un metal divalente M^{2+}, el cual se encuentra tetraédricamente coordinado a través de una interacción ion dipolo con cuatro residuos conservados, los cuales son: Asp242, His246, His330, and His357, tomando como base de numeración la secuencia humana, dichos aminoácidos se muestran en la figura 2 y 7 (sección de resultados). Un aspecto de llamar la atención son las enzimas de la subfamilia MAR (cd08177), las cuales presentan actividad pese a carecer de iones metálicos en el centro activo. En estas proteínas el Asp242 se encuentra sustituido por una asparagina o una arginina, como puede observarse en la figura 2 y 6 (sección de resultados). Aquí vale la pena aclarar, que los miembros de la subfamilia FeADH2 (cd08183), de la cual ninguno ha sido caracterizado aun, así como miembros de la subfamilia HEPD (cd08182), descritas como hidroxetilfosfato deshidrogenasa o fosfonoacetaldehído reductasa, probablemente sean también funcionales en ausencia de iones divalentes, esto debido a que el Asp242 se encuentra remplazado por una glutamina en dicha posición, como se muestra en la figura 2 (sección de resultados).

Adicionalmente, en estudios previos usando mutagénesis sitio-dirigidas, se propuso la participación de la His267 en la unión al Fe^{2+} esto en la enzima FucO de *E. coli* (presentando un residuo de Tyr334) de acuerdo a la ADHFe1 en humanos. Sin embargo, la estructura cristalina de FucO de *E. coli* mostró que dicha His267 no se encuentra coordinada con el ión Fe^{+}. Recientemente, estudios hechos por Fujii y cols. (Fujii *et al*., 2016) donde compararon las estructuras de

FucO de *E. coli* y GraC de *Rhizobium sp.* MTP-10005, esta última con actividad de maleíl acetato reductasa propusieron que la His267 de FucO corresponde a la His243 de GraC, donde ambos residuos podrían interactuar con el sustrato y por lo tanto deberían estar participando en la catálisis más que en la unión a metal.

Es importante mencionar que a pesar de que los miembros de la familia ADHFe son descritos generalmente como alcohol deshidrogenasas activadas por hierro, existe, en realidad, una gran variedad de cationes divalentes que pueden actuar como activadores como son: el zinc, níquel, magnesio, cobre, cobalto o manganeso. Inclusive dentro de las enzimas activadas por hierro como FucO de *E. coli* o la ADHII de *Z. mobilis*, ambas pertenecientes a la subfamilia LPO, en donde el hierro puede ser reemplazado por zinc. Más específicamente en FucO de *E. coli* donde la enzima es activada únicamente por Fe^{2+}, y donde el zinc inactiva la enzima, FucO presenta una más alta afinidad por el $Zinc^{2+}$ en comparación con el Fe^{2+}.

En el caso de la familia de glicerol deshidrogenasa, el ion metálico, presente en dicho grupo de enzimas corresponde al átomo de zinc, el cual se encuentra coordinado por dos residuos de histidina y uno de aspartato. Por lo que únicamente un residuo de histidina se encuentra conservado en ambas familias, correspondiente a la posición 357, tomando como referencia la ADHFE1 de ser humano, cabe mencionar que dicha histidina es usada en ambas familias para coordinar ya sea un átomo de hierro en la familia ADHFe o un átomo de zinc en la familia glicerol deshidrogenasa, como se observa en la figura 2 (sección de resultados). La existencia de distintos residuos de unión a metal entre estas dos familias, ADHFe y glicerol deshidrogenasa, reafirma a este último grupo de enzimas como parte de otra familia de proteínas aunque estrechamente relacionadas.

6.6 Principales subfamilias de ADHFe presentes en eucariontes

Todas las secuencias de ADHFe de eucariontes se agrupan en 13 diferentes subfamilias proteicas pertenecientes a la familia de las ADHFe. Dichas secuencias no se encuentran distribuidas homogéneamente en estas trece subfamilias, ya que tan solo cuatro de ellas comprenden poco más del 92% del total de las secuencias a la fecha recuperadas para el presente estudio.

6.6.1 Subfamilia hidroxiácido oxoácido transhidrogenasa (cd08190)

Subfamilia HOT (cd08190): algunas proteínas de esta subfamilia han sido caracterizadas en mamíferos y poseen actividad de hidroxiácido oxoácido transhidrogenasa HOT, que cataliza la conversión de ácido gamma hidroxibutírico a succinato semialdehído. Cabe mencionar que pese a que este tipo de enzimas presentan un motivo de unión a dinucleótido, la reacción catalizada por dicha enzima no es la de reducir a la coenzima, más bien se encuentra acoplada a la reducción del alfa cetoglutarato. En seres humanos el gen que codifica a la proteína fue denominado ADHFE1 por el comité de nomenclatura genómica, por sus siglas en inglés HUGO. El sustrato de la enzima, el ácido gamma hidroxibutírico (GHB) es un compuesto presente de manera natural en el cerebro y tejidos periféricos de los mamíferos. El GHB es de interés debido a que es un compuesto natural con propiedades neuromodulatorias en las sinapsis GABAergicas. Por otro lado ha sido reportado que el GHB actúa como un regulador energético, de tal manera que algunas atletas lo usan ya que ha sido demostrado que elevan los niveles de hormona del crecimiento *in vivo* en seres humanos. Un estudio encontró que el GHB duplicó la secreción de hormona de crecimiento en hombres jóvenes. El aumento de los niveles de hormona de crecimiento por GHB es mediado a través de receptores muscarínicos de acetilcolina, como se demostró por la administración de pirenzepina, un agente que bloquea dichos receptores. Dado que ha sido demostrado que ciertas sales de

succinato elevan los niveles de hormona de crecimiento *in vitro*, y debido a que el GHB es metabolizado a succinato semialdehído, algunas personas han sugerido que este último juega un papel en el aumento de los niveles de hormona del crecimiento provocado por el GHB. Sin embargo, no existe evidencia que demuestre que el succinato juegue algún papel en aumentar los niveles de hormona del crecimiento por parte del GHB (Deng *et al*., 2002; Kardon *et al*., 2006; Kaufman *et al*., 1988; Kaufman *et al*., 1991; Koch *et al*., 2002 Lyon *et al*., 2009; Nelson *et al*., 1981; Maitre, *et al*., 2016; Van Cauter *et al*., 1997; Volpi *et al*., 1997; Volpi *et al*., 2000).

Adicionalmente, existe evidencia de que el metabolismo endógeno de GHB parece estar asociado con la habilidad natural atlética. Esta idea tiene como respaldo datos que identifican a ADHFE1 como un gen candidato a mejorar el desempeño atlético, el cual ha sido blanco de selección positiva durante 400 años en caballos pura sangre (McGivney *et al*., 2009; Gu *et al*., 2009).

El gen ADHFE1 es expresado principalmente en adultos en tejidos como hígado, riñón, corazón y adipocitos, en el hipotálamo y células del neuroblastoma, además de diversos tejidos fetales. En contraste, transcritos del gen ADHFE1 no son detectados en pulmón intestino, estómago, túbulos seminíferos, músculo y testículos. Por otro lado Tae y colaboradores (Tae *et al*., 2013) demostraron que la expresión del gen ADHFE1 es mayor en tejidos bien diferenciados que aquellos en proceso de diferenciación, mientras que líneas celulares de cáncer colorectal mostraron una regulación a la baja de los mRNAs del gen ADHFE1 y por ende su proteína debido a hipermetilación del promotor de ADHFE1. Por lo tanto, se puede deducir que ADHFE1 tiene un importante papel en órganos con alta actividad metabólica, así como en la diferenciación y en el proceso de desarrollo embrionario.

En cuanto a la localización celular, existe evidencia de pruebas inmunocitoquímicas que revelan una localización mitocondrial para la ADHFE1 en ratones. Adicionalmente, nosotros realizamos pruebas de predicción celular con diferentes softwares de uso libre como son: MITOPRED, WoLF PSORT y PredSL.

Dichas pruebas sugieren con un alto valor de probabilidad que todas las ADHFE1 son mitocondriales. Estas enzimas conservan tanto el sitio de unión a coenzima como el motivo de unión a hierro, como se muestra en la figura 2 y 6 de la sección de resultados. Este tipo de enzimas presentan un nucleótido fuertemente unido, por lo que no requieren la adición de NAD(P) para desplegar actividad catalítica. Debido a que la oxidación del etanol esta acoplada a la reducción de un cofactor como NAD(P), la participación de la ADHFe es cuestionable (Guda *et al*., 2004; Horton *et al*., 2007; Petsalaki *et al*., 2006; Kim *et al*., 2007). Además, la exposición a etanol en adipocitos humanos en concentraciones que van de 1 a 100 mM, no modifica los niveles de transcripción de ADHFE1, reforzando que esta enzima no participa en el metabolismo del etanol en animales.

Todas las ADHFE1 en animales corresponden a genes de copia única, a pesar de que se ha reportado que el genoma completo de los vertebrados ha sufrido varias duplicaciones a lo largo de su historia evolutiva. Dado que todos los animales tienen una copia del gen ADHFE1, la cual pertenece a la familia HOT, se puede deducir que esta proteína está realizando actividades esenciales en animales.

De acuerdo a nuestro alineamiento estructural, logramos identificar que todas las secuencias pertenecientes a la subfamilia HOT poseen dos inserciones. La primera de estas inserciones corresponde a un fragmento de 19 residuos aminoácidos, residuos del 256 al 274 de acuerdo a la ADHFE1 en humano, dicho inserto se encuentra ubicado entre las hélices alfa número 5 y 6; el segundo inserto corresponde a un fragmento de 13 residuos aminoácidos, residuos de 342 al 354 de acuerdo a la ADHFE1 en seres humanos, dicho inserto se encuentra ubicado entre las hélices alfa número 8 y 9, como se observa en las figuras 7, 8 y 9 (sección de resultados). Vale la pena mencionar que ambos insertos se encuentran ausentes en todos los otros miembros de la familia ADHFe incluso de la familia proteica de las glicerol deshidrogenasa, siendo exclusivos de la subfamilia HOT.

Dado que miembros de la subfamilia HOT son encontrados en los tres dominios de la vida, este grupo de enzimas es probablemente uno de los más antiguos

dentro de la familia ADHFe. A pesar de esto, la actividad realizada por estas enzimas en el resto de los organismos, no animales, es desconocida.

De igual forma se realizó un análisis de expresión del gen adhfe1 con ayuda de GEO Profiles en la página del NCBI, corroborando lo que ya se había reportado antes, que los niveles de expresión de tal gen no cambian en los distintos órganos por la administración de etanol (Cuadro 3 de la sección de resultados).

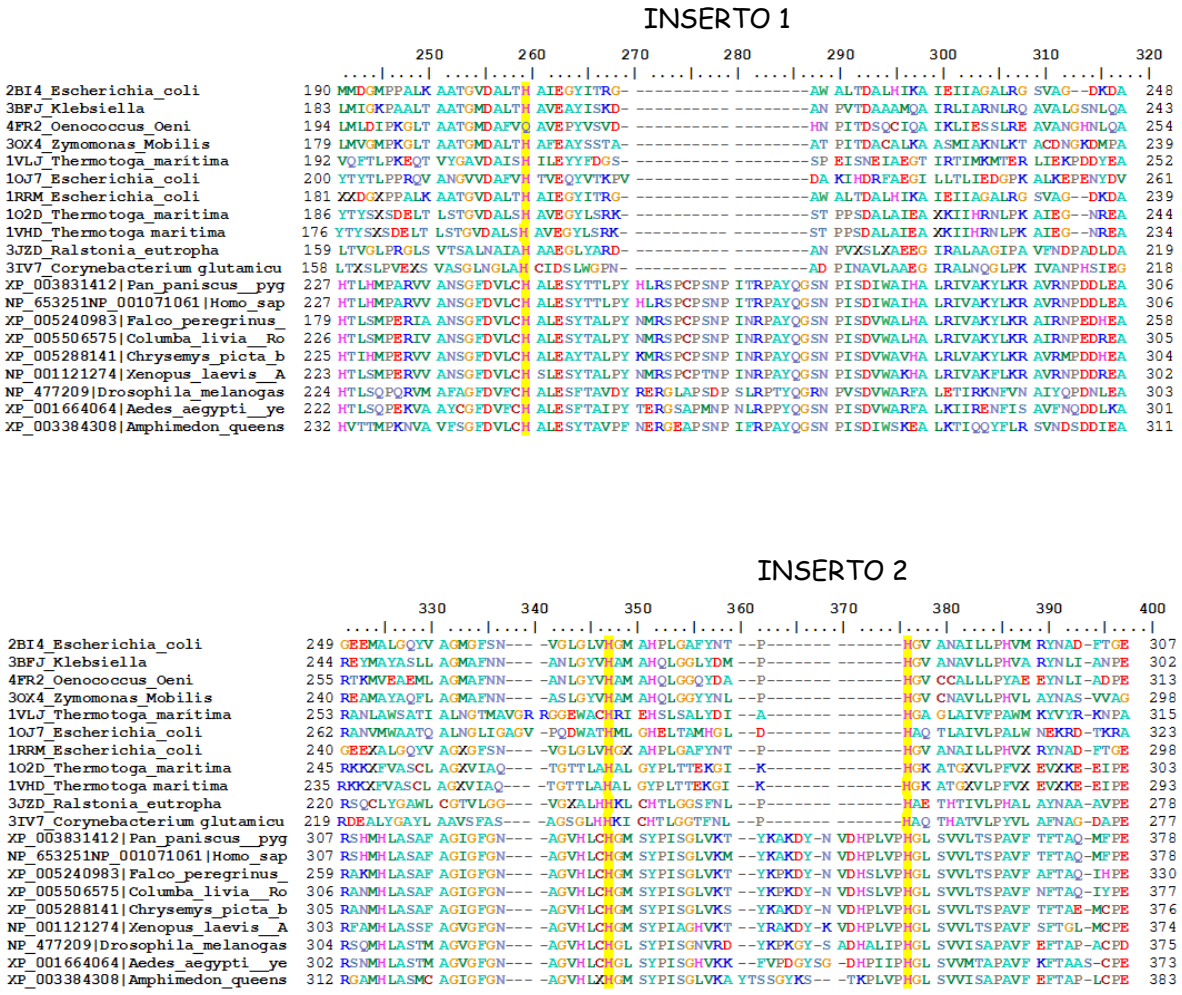


Figura 7. Ubicación de los dos insertos posibles responsables del cambio de actividad enzimática en las enzimas de la subfamilia HOT, dichos insertos se encuentran únicamente en las secuencias de animales y una rama del grupo hermano Fungi. En amarillo se resaltan los residuos de Histidina que forman enlaces de coordinación con el Fe.

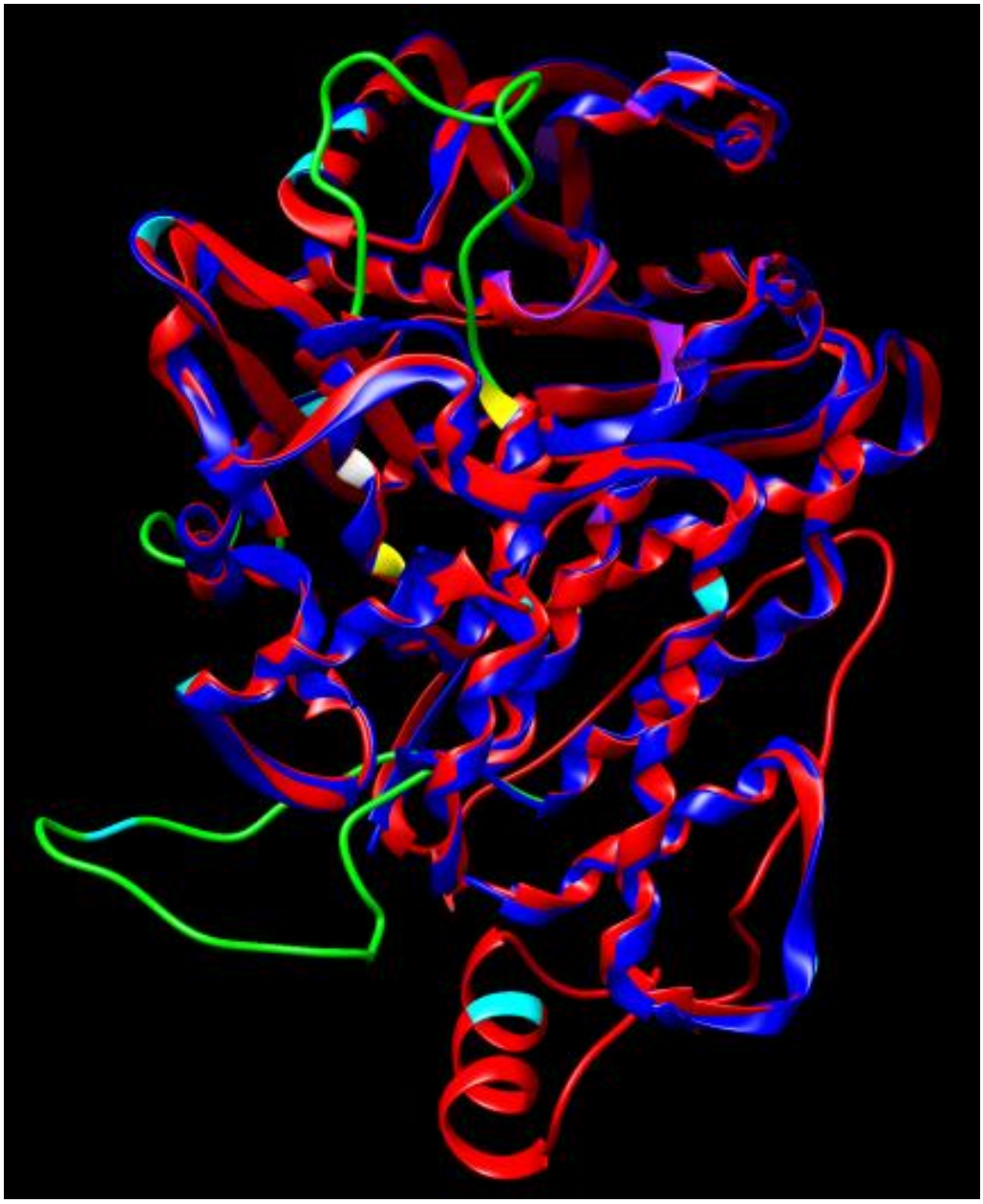


Figura 8. Modelo de la ADH-Fe humana. Se utilizó el sofware de I-TASSER para predecir la estructura terciaria de la ADH-Fe humana. Dicho software detecta estructuras que usa como templado de la base de datos del PDB, mediante la técnica de reconocimiento de plegamiento (fold recognition o threading). En rojo esta la secuencia de humano y en azul oscuro la secuencia con número de acceso 3BFJ (propanodiol deshidrogenasa) de *Klebsiella*. En verde se resaltan los dos insertos identificados en nuestro análisis estructural para el grupo de enzimas que presentan actividad Hidroxiácido-Oxoácido Transhidrogenasa (HOT). En azul claro se muestran los aminoácidos sujetos a selección positiva; en morado aquellos aminoácidos que interaccionan con NAD^+ o $NADP^+$ y en amarillo las Histidinas que forman enlace de coordinación con el átomo de Fe^+.

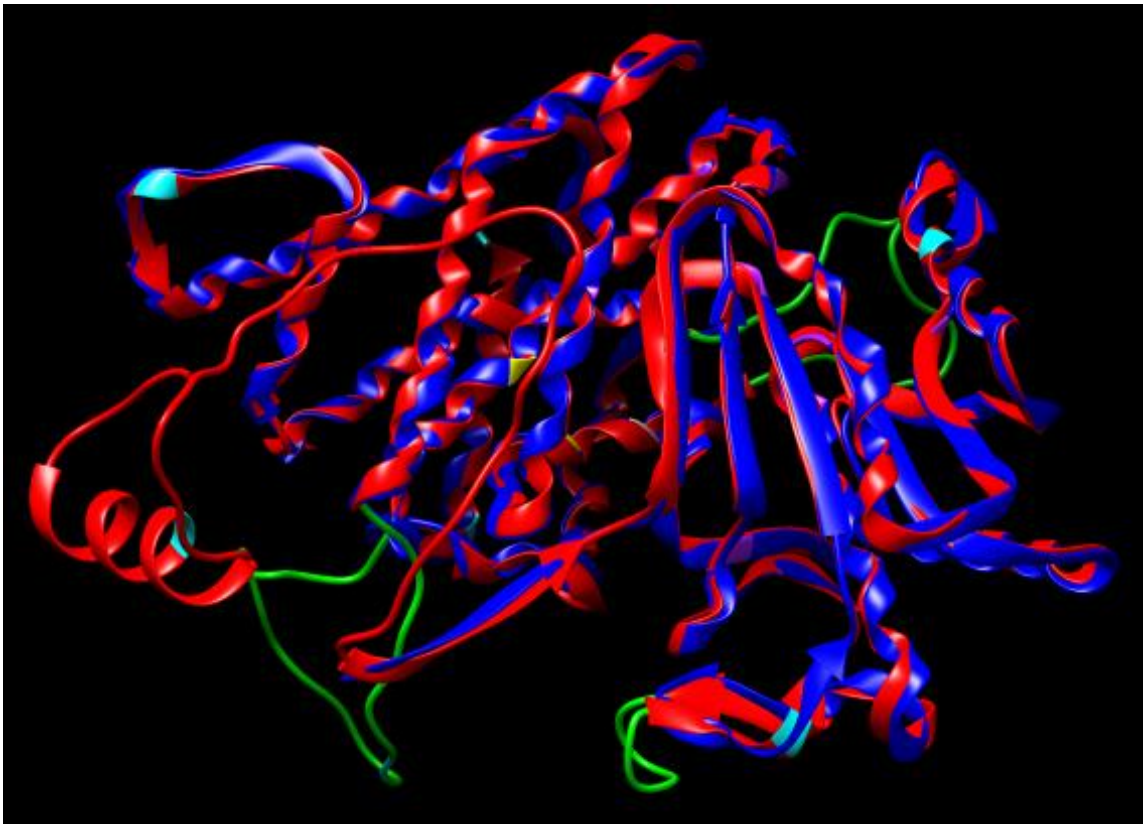


Figura 9. Modelo de la ADH-Fe humana. Modelado mediante el software I-TASSER. En verde se resaltan los dos insertos identificados en nuestro análisis estructural para el grupo de enzimas que presentan actividad Hidroxiácido-Oxoácido Transhidrogenasa (HOT), en amarillo se resaltan las Histidinas que forman enlace de coordinación con el Fe+, una de las cuales está al lado de un inserto.

Cuadro 3. Efecto de la administración de etanol sobre la expresión del gen adhfe en distintos tejidos y animales. Realizado a partir de la base de datos de GEO Profiles en la página del NCBI.

Experimental model	specie	Organ/Tissue /cells	Effect on gene expression	DataSet accession[1]	Reference
In vivo models					
Fetal alcohol spectrum disorder model Embryos treated with 100 mM ethanol in embryo medium from 2 to 8 hours post fertilization.	*Danio rerio*	Whole embryo	No change	GDS5082	Sarmah et al., (2013)
Effect of long-term ethanol consumption Animals were fed with ethanol containing Lieber-DeCarli diet for 8 weeks.	*Rattus norvegicus* (Wistar males)	pancreas	No change	GDS2107	Kubisch et al., (2006)
Offspring exposed to ethanol during neurodevelopment by maternal continuous preference drinking. Brains from adult male offspring prenatally exposed to alcohol via voluntary maternal consumption (10% ethanol in drinking water) from pregestational day -14 to pup postnatal day 10.	*Mus musculus*	70 days postpartum male offspring (whole brain tissue)	No change	GDS4613	Kleiber et al., (2012)
Time course effect of addictive drugs on brain striatum Animals were treated for up to 8 hours with ethanol (2 g/kg i.p.).	*Mus musculus* (C57BL/6J mice)	brain striata	No change at any time	GDS3703	Piechota et al., (2010); Korostyynski et al., (2013)
In vitro models					
Effect of ethanol exposure in adipocytes Cells were exposed to 1 mM through 100 mM of ethanol for 6, 12, 24, 48, or 72 h.	*Mus musculus*	Adipocytes (3T3-L1 cells)	No change at any time or dose	--	Kim et al., (2007)

[1]All public microarray gene expression datasets were downloaded from NCBI's Gene Expression Omnibus (GEO) http://www.ncbi.nlm.nih.gov/geoprofiles

6.6.2 Subfamilia Lactaldehído propanodiol oxidoreductasa (cd08176)

Esta subfamilia proteica incluye proteínas con diferentes actividades catalíticas, como se observa en el cuadro 1 de la sección de resultados. Entre las actividades reportadas podemos encontrar: lactaldehído: propanodiol oxidoreductasa (lactaldehído reductasa), L-1,3-propanodiol deshidrogenasa, metanol deshidrogenasa, alcohol deshidrogenasa, metanol deshidrogenasa, alcohol deshidrogenasa y L-treonina deshidrogenasa. En eucariontes únicamente la ADH4 de *Saccharomyces cerevisiae* ha sido caracterizada completamente como miembro de la subfamilia LPO. *Saccharomyces cerevisiae* posee 5 diferentes isoenzimas de alcohol deshidrogenasa (Adh). Estudios hechos con cepas cultivadas con glucosa o etanol como fuente de carbono revelaron que la ADH1

era la única alcohol deshidrogenasa capaz de catalizar eficientemente la reducción de acetaldehído a etanol. Una cepa mutante con el gen ADH4 intacto fue capaz de crecer en glucosa pero a una tasa mucho más lenta que la cepa silvestre, produciendo incluso menos etanol a partir de la glucosa, siendo también incapaces de utilizar el etanol como fuente de carbono. En contraste, altos niveles de glicerol y acetaldehído fueron observados en dicha mutante. Debido a que no se observa la transcripción del gen ADH4 en cepas que crecen en etanol y cepas con el gen ADH4 intacto fueron incapaces de crecer en etanol, es muy probable que la expresión de dicho gen no se encuentre relacionada al consumo de etanol, a pesar de que las propiedades cinéticas de ADH4 en comparación a las otras isoenzimas, mostró que el etanol es un sustrato adecuado para la ADH4. Teniendo como mejores sustratos al etanol y al n-propanol. Así aunque las propiedades cinéticas de la ADH4 hacen de esta enzima, un candidato para el metabolismo del etanol, es posible que esta enzima desarrolle papeles fisiológicos adicionales.

6.6.3 Subfamilia de doble dominio alcohol acetaldehído deshidrogenasa (cd08178)

Miembros de esta subfamilia también han sido reportados en bacterias, organismos pertenecientes al reino Fungi, al grupo de las clorofitas y varios eucariontes inferiores. Estas enzimas con doble dominio y función, son también conocidas como ADHEs y se encuentran presente en una variedad de organismos fermentativos. Estas enzimas catalizan la conversión de acil-coenzima A a alcohol vía un aldehído intermedio; tal reacción se encuentra acoplada a la oxidación de dos moléculas de NADH, con la intención de mantener la poza de NAD^+ durante el metabolismo fermentativo. ADHE forma un gran ensamble multimérico helicoidal también llamado espirosoma, que consiste de un dominio aldehído deshidrogenasa N-terminal, el cual pertenece a la familia proteica de las ALDH20, y un dominio C-terminal con actividad alcohol deshidrogenasa (ADH) que es miembro de la subfamilia AAD-C.

Es generalmente aceptado que las ADHEs realizan la importante función de regenerar NAD^+ a partir de NADH en condiciones anaeróbicas para mantener en marcha la glucólisis. Se ha observado que cepas de la bacteria *Entamoeba histolytica* con ablación del gen ADHE, presentan una total inhibición en la producción de etanol, además, de la incapacidad de dichas cepas de sobrevivir en condiciones anaerobias, juntas ambas evidencias sugieren fuertemente que la ADHE cumple un papel en la producción de etanol.

En *Escherichia coli* donde ambas rutas glicolítica y el ciclo de los ácidos tricarboxílicos son las dos más importantes fuentes de intermediarios metabólicos y de cofactores reducido NADH. Sin la re oxidación del NADH la poza de NAD^+ se vería rápidamente mermada, deteniendo el metabolismo celular y el crecimiento. En estos casos cuando el oxígeno está presente, los equivalentes reducidos de NADH son transferidos a la cadena de transporte de electrones (ETC), dando como producto final H_2O y un potencial de membrana que es usado para sintetizar ATP. Esto regenera NAD^+ para luego ser usado en reacciones subsecuentes. Cuando la enterobacteria crece anaeróbicamente, un sistema de transporte de electrones está disponible solo si alterna aceptores de electrones, tales como nitrato, N-oxido trimetilamina o fumarato están presentes (Nimmo, 1987; Poole y Ingeldew, 1987).

Si los oxidantes alternativos están ausentes, la bacteria debe reciclar el NADH de alguna otra manera, es decir mediante la fermentación. Esto involucra el uso de NADH para reducir intermedios metabólicos y excretar los productos finales dentro del medio de crecimiento. Este proceso resulta en la pérdida de la mayoría de la energía contenida disponible en las fuentes de carbono pero permite a la bacteria crecer, aunque sea a una tasa reducida (Ingeldew y Poole, 1984).

Adicionalmente Atteia y cols. (Atteia *et al*., 2013) reportaron la presencia de una ADHE en mitocondria aislada de la clorofita *Polytonella sp*. La co expresión tanto en ambientes aerobios como anaerobios de la ADHE sugiere que esta enzima podría estar involucrada ya sea en el mantenimiento del balance redox (producción de etanol) o en la asimilación del etanol (produciendo acetil-CoA y

NADH por medio de la respiración), o ambos dependiendo de las condiciones ambientales.

Finalmente, es interesante mencionar que la ADHE de *E. coli* puede unirse al ribosoma 70S bacteriano, exhibiendo actividad a la hora de desenrollar el RNA. De esta manera la ADHE puede funcionar en *E. coli* al menos, como una proteína regulatoria, revelando una inesperada acción moonlighting, dejando la posibilidad de que otras ADHE presenten funciones adicionales.

6.6.4 Subfamilia de Maleíl-acetato reductasa (cd08177)

Las proteínas de esta subfamilia, son enzimas clave para la degradación de los productos del rompimiento del anillo derivados de la degradación aeróbica de compuestos aromáticos de microorganismos. Estas enzimas catalizan la reducción del Maleíl acetato a partir de NADH o NADPH, sobre un doble enlace C-C para obtener 3-oxoadipato. Proteínas homólogas de MAR son encontradas en los tres dominios de la vida. En eucariontes, proteínas de esta subfamilia están presentes principalmente en el reino fungi, incluyendo tanto Basidiomicetos como Ascomicetos. Mientras que en eucariontes inferiores homólogos de MAR fueron encontrados en *Emiliania huxleyi* del grupo Haptophyceae y en *Nannochloropsis gaditana* del grupo Stramenopiles. En organismos pertenecientes al reino Fungi, las maleíl acetato reductasas contribuyen al catabolismo de una variedad de sustratos relacionados, tal como tirosina, resorcinol, fenol, hidroquinona, gentisato, benzoato, 4-hidroxibenzoato, protocatechuate, vainillato e inclusive contaminantes aromáticos. En *Fusarium verticilloides* un gen homólogo de MAR identificado como FUM7 fue encontrado agrupado con otros genes involucrados en la biosíntesis de fumonisina.

La fumonisina son mico toxinas policétidas que pueden acumularse en plantas infectadas con un hongo y causar varias enfermedades fatales en animales, como son; leucoencefalomalacia en caballos, edema pulmonar en cerdos, cáncer en ratas y ratones y cáncer de esófago en humanos.

Mutantes con el gen *FUM7* suprimido producen análogos de la fumonisina con función alqueno. Esto sugiere que FUN7 probablemente catalice la reducción de un alqueno intermediario de la biosíntesis de fumonisina, en una reacción similar a aquella realizada por las maleíl acetato reductasas.

6.7 Resultados Adicionales

Como ya habíamos reportado antes, mediante nuestro alineamiento estructural, identificamos dos insertos los cuales solo están presentes en las secuencias de animales, de algunos hongos, así como unas pocas bacterias y protistas, lado izquierdo del árbol, correspondiente a la subfamilia HOT (figura 4 de la sección de resultados). Dichos insertos se encuentran formando asas (figura 8) según el modelo de predicción de estructura tridimensional I-TASSER, para la secuencia de ADH-Fe humana y una de estas asas colinda con una histidina que forma un enlace de coordinación con el átomo de Fe (Figura 7 y 8 de la sección de resultados). Como se observa en la figura 10 (sección de resultados), se identificaron los codones sujetos a selección positiva (9 codones), mediante el software DataMonkey. Dicho análisis hecho mediante el modelo de evolución de efectos mezclados (MEME) para ambas secuencias animales y hongos, arrojó que existen 9 codones sujetos a selección positiva. Dichos codones codifican para los aminoácidos A59, E63, K69, V92, H204, R392, Y467; tomando como referencia la secuencia proteica de humano. Luego de haber identificado los codones sujetos a selección positiva y ubicarlos en la secuencia proteica procedimos al análisis de Logos (Figura 10), donde se observa que únicamente tres de los siete aminoácidos se encuentran bien conservados. Cabe mencionar que el análisis de MEME se hizo en 4 ocasiones, modificando parámetros como son la longitud de las secuencias y el número de éstas. Para el análisis de Logos así como los aminoácidos esquematizados en la figura 10 (sección de resultados), sólo se usaron los codones que fueron seleccionados por estar bajo selección positiva en todos los análisis. En el caso de los aminoácidos no conservados, identificamos que los posibles cambios entre las distintas secuencias involucrarían mutaciones en dos pares de bases en cada codón.

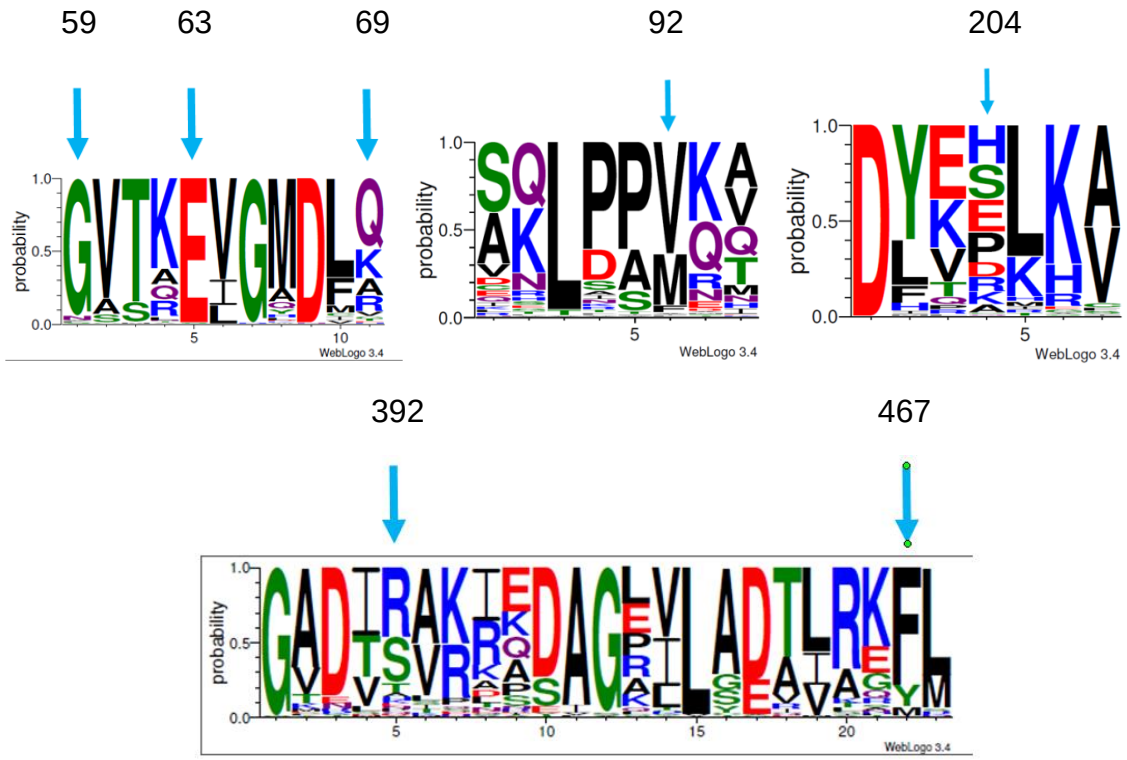


Figura 10. Análisis de Logos donde se aprecia, que tan bien conservados están los aminoácidos codificantes por los codones sujetos a selección positiva. Se tomó como base de la numeración la secuencia de ADH-Fe humana.

7. Conclusiones

Las ADHFe pertenecen a una antigua familia proteica que puede ser encontrada en los tres dominios de la vida. Dicha familia está formada por al menos 19 diferentes subfamilias con una gran diversidad de enzimas que desarrollan varias funciones metabólicas. Muchas ADHFe son activadas o contienen un átomo de hierro, pero existen también enzimas que presentan algún otro ion divalente, como el zinc o incluso pueden carecer de cofactor metálico. En eucariontes la mayoría

de las ADHFe pertenecen a la subfamilia HOT con número cd08190. Más específicamente, el 100% de dichas secuencias encontradas en animales y el 80% de las secuencias de hongos pertenecen a esta subfamilia HOT. Las plantas carecen de proteínas clasificadas en la subfamilia HOT. El resto de las ADHFe de eucariontes muestra una distribución filética irregular y son clasificadas en doce subfamilias proteicas adicionales, siendo la más importante la subfamilia maleíl-acetato reductasa (MAR) encontrada principalmente en organismos del reino Fungi, por último la subfamilia de las lactaldehído: propanodiol deshidrogenasa (LPO) y la subfamilia de doble dominio aldehído deshidrogenasa y alcohol deshidrogenasa (AAD) se encuentran en fungi, clorofitas y eucariontes inferiores.

8. Referencias

1. Akashi, H. 2012. Weak Selection and Protein Evolution. Genetics. 192 (1): 15–31.
2. Almeida P. 2016. Proteins. Concepts in Biochemistry. Garland Science, Taylor & Francis Group, LLC.
3. Altman R. B. 2004. Building successful biological databases. Brief. Bioinformatics 5 (1): 4-5.
4. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. J Mol Biol, 215, 403–410.
5. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 25, 3389–3402.
6. Andersson J O, Hirt R P, Foster P G, Roger A J. 2006. Evolution of four gene families with patchy phylogenetic distributions: influx of genes into protist genomes. BMC Evol Biol 6: 27. doi: 10.1186/1471-2148-6-27 PMID: 16551352
7. Andreeva, A., Howorth, D., Chandonia, J.-M., Brenner, S.E., Hubbard, T.J.P., Chothia, C., and Murzin, A.G. 2008. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res, 36, D419–D425.
8. Aravind L, Anantharaman V, Koonin E. V. 2002. Monophyly of class I aminoacyl tRNA synthetase, USPA, ETFP, photolyase, and PP-ATPase nucleotide-binding domains: implications for protein evolution in the RNA. Proteins 48:1-14.
9. Arfman N, Hektor HJ, Bystrykh LV, Govorukhina NI, Dijkhuizen L, Frank J. 1997. Properties of an NAD(H)-containing methanol dehydrogenase and its activator protein from Bacillus methanolicus. European Journal of Biochemistry 244: 426-433.
10. Attwood T. K., Gisel A., Eriksson N-E. and Bongcam-Rudloff E. 2011. Concepts, Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective. Bioinformatics - Trends and Methodologies. InTech.

11. Bailey T. L., Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology. Menlo Park, California: AAAI Press. pp. 28-36.
12. Bailey T. L., Gribskov M. 1998. Combining evidence using p-values: application to, sequence homology searches. Bioinformatics 14 (1): 48-54.
13. Baier F, Copp J.N, Tokuriki N. 2016. Evolution of enzyme superfamilies: comprehensive exploration of sequence-function relationships. Biochemistry. American Chemical Society.
14. Bashton M, Chothia C. 2007. The generation of new protein functions by the combination of domains. Structure 15:85-99.
15. Bashton M, Nobeli I, Thornton J.M. 2006. Cognate ligand domain mapping for enzymes. J. Mol. Biol. 364:836–852.
16. Battelli F, Stern L. 1909. L'alcoolase dans les tissus animaux. Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences, Paris 67: 419-421.
17. Battelli F, Stern L. 1910. Die Alkoholoxydase in den Tiergeweben. Biochem Z 28: 145-168.
18. Bateman, A., Coggill, P., and Finn, R.D. 2010. DUFs: families in search of function. Acta Crystallogr, 66, 1148–1152.
19. Bolser D. M., Chibon P. Y., Palopoli N; 2012. MetaBase-- the wiki-database of biological databases. Nucleic. Acids. Res. 40 (Database issue): D1250-1254.
20. Boratyn G.M, Camacho C, Cooper P.S, Coulouris G, Fong A, Ma N, Madden T.L, Matten W.T, McGinnis S.D, Merezhuk Y, Raytselis Y, Sayers EW, Tao T, Ye J, & Zaretskaya I 2013. BLAST: a more efficient report with usability improvements. Nucleic Acids Res., 41, W29-W33.
21. Bourne P. 2005. Will a biological database be different from a biological journal? PLoS Comput. Biol. 1 (3): 179-81.
22. Bowie J.U., Luthy R, Eisenberg D; Lüthy; Eisenberg 1991. A method to identify protein sequences that fold into a known three-dimensional structure. Science 253 (5016): 164-170.

23. Brazas M.D; Yim D. S; Yamada J.T; Ouellette B.F. 2011. The 2011 Bioinformatics Links Directory update: more resources, tools and databases and features to empower the bioinformatics community. Nucleic. Acids. Res. 39 (Web Server issue): W3–7.
24. Björklund, A. K. *et* al. 2006. Expansion of Protein Domain Repeats. PLoS Comput. Biol., 2(8), e114.
25. Björklund A.K. 2005. Domain rearrangements in protein evolution. J. Mol. Biol. 353(4): 911-923.
26. Brylinski M. 2013. The utility of artificially evolved sequences in protein threading and fold recognition. J. Theor. Biol. 328: 77-88.
27. Budd, Aidan 2009. Multiple sequence alignment exercises and demonstrations. European Molecular Biology Laboratory.
28. Carter S. O. and Gibson M. 2005. γ-Hydroxybutyric Acid. The New England journal of medicine 352:2721-2732.
29. Chen C, Natale D.A, Finn R.D, Huang H, Zhang J, Wu C.H, & Mazumder R. 2011. Representative proteomes: a stable, scalable and unbiased proteome set for sequence analysis and functional annotation. PLoS One., 6, e18910.
30. Choudhuri. S. 2014. BIOINFORMATICS FOR BEGINNERS: Genes, Genomes, Molecular Evolution, Databases and Analytical Tools. Published by Elsevier Inc.
31. Chothia, C. and Lesk, A.M. 1986. The relation between the divergence of sequence and structure in proteins. EMBO J, 5, 823–826.
32. Chothia, C., Gough, J., Vogel, C., and Teichmann, S.A. 2003. Evolution of the Protein Repertoire. Science, 300, 1701-1703.
33. Chun S, Fay J. C. 2011. Evidence for hitchhiking of deleterious mutations within the human genome. PLoS Genet 7:e1002240.
34. Conway T, Ingram L.O. 1989. Similarity of Escherichia coli propanediol oxidoreductase (fucO product) and an unusual alcohol dehydrogenase from *Zymomonas mobilis* and *Saccharomyces cerevisiae*. J Bacteriol 171: 3754-3759.

35. Costantini S, Colonna G, Facchiano A.M. 2006. Amino acid propensities for secondary structures are influenced by the protein structural class. Biochem Biophys. Res. Commun. 342(2):441-51
36. Csaba, G., Birzele, F., and Zimmer, R. 2009. Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis. BMC Struct Biol, 17(9), 23.
37. Danielsson O, Jörnvall H. 1992. Enzymogenesis: classical liver alcohol dehydrogenase origin from the glutathione-dependent formaldehyde dehydrogenase line. Proceedings of the National Academy of Sciences of the United States of America. 89 (19): 9247-51.
38. Daubaras D.L, Saido K, Chakrabarty AM. 1996. Purification of hydroxyquinol 1,2-dioxygenase and maleylacetate reductase: the lower pathway of 2,4,5-trichlorophenoxyacetic acid metabolism by Burkholderia cepacia AC1100. Appl Environ Microbiol 62: 4276-4279.
39. Dayhoff, M. O.; McLaughlin, P. J.; Barker, W. C.; Hunt, L. T. 1975. Evolution of sequences within protein superfamilies. Die Naturwissenschaften. 62 (4): 154–161.
40. Dayhoff, M. O. 1976. The origin and evolution of protein superfamilies. Federation proceedings. 35 (10): 2132–2138.
41. de Vries GE, Arfman N, Terpstra P, Dijkhuizen L. 1992. Cloning, expression, and sequence analysis of the Bacillus methanolicus C1 methanol dehydrogenase gene. J Bacteriol 174: 5346-5353.
42. Dereeper A., Guignon V., Blanc G., Audic S., Buffet S., Chevenet F., Dufayard J.F., Guindon S., Lefort V., Lescot M., Claverie J.M., Gascuel O. 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 1:36 (Web Server issue):W465-9.
43. Deng Y, Wang Z, Gu S, Ji C, Ying K, Xie Y, & Mao Y. 2002. Cloning and characterization of a novel human alcohol dehydrogenase gene (ADHFe1). DNA Seq., 13, 301-306.
44. Dor O, Zhou Y; Zhou. 2006. Achieving 80% tenfold cross-validated accuracy for secondary structure prediction by large-scale training. Proteins 66 (4): 838-845.

45. Duester G. 2008. Retinoic acid synthesis and signaling during early organogenesis. Cell. 134 (6): 921-31.
46. Dunbrack R. L. 2006. Sequence comparison and protein structure prediction. Curr. Opin. Struct. Biol. 16 (3): 374-84.
47. Duret, L. 2008. Neutral theory: The null hypothesis of molecular evolution. Nature Education. 1 (1): 803–6. 218.
48. Farrés J, Moreno A, Crosas B, Peralba J.M, Allali-Hassani A, Hjelmqvist L, Jörnvall H, Parés X. 1994. Alcohol dehydrogenase of class IV (sigma sigma-ADH) from human stomach. cDNA sequence and structure/function relationships. European Journal of Biochemistry / FEBS. 224 (2): 549–57.
49. Farris, J. S. 1970. Methods for computing Wagner trees. Systematic Zoology 19, 83-92.
50. Felsenstein J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783-791.
51. Felsenstein J. 2004. Inferring Phylogenies Sinauer Associates: Sunderland, Massachusetts.
52. Finn R.D, Bateman A, Clements J, Coggill P, Eberhardt R.Y, Eddy S.R, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, & Punta M. 2014. Pfam: the protein families database. Nucleic Acids Res., 42, D222-D230.
53. Finn R.D, Coggill P, Eberhardt R.Y, Eddy S.R, Mistry J, Mitchell A.L, Potter S.C, Punta M, Qureshi M, Sangrador-Vegas A, Salazar G.A, Tate J, & Bateman A. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res., 44, D279-D285.
54. Fitch, W. M. 1971. Toward defining the course of evolution: minimum change for a specified tree topology. Systematic Zoology 20 (4), 406-416
55. Fitch W. M, Margoliash E. 1967. Construction of phylogenetic trees. Science 155: 279-84.
56. Fujii T, Sato A, Okamoto Y, Yamauchi T, Kato S, Yoshida M, 2016. The crystal structure of maleylacetate reductase from Rhizobium sp. strain MTP-10005 provides insights into the reaction mechanism of enzymes in its original family. Proteins 2016; 84:1029-1042.

57. Gascuel O, Steel M. 2006. Neighbor-joining revealed. Mol. Biol. Evol. 23 (11): 1997-2000.
58. Gandhimathi, A.; Nair, A. G.; Sowdhamini, R. 2011. PASS2 version 4: An update to the database of structure-based sequence alignments of structural domain superfamilies. Nucleic Acids Research. 40 (Database issue): D531-D534.
59. Glasner M.E., Gerlt J.A., Babbitt P.C. 2006. Evolution of enzyme superfamilies. Curr. Opin. Chem. Biol. 10:492-497.
60. Goncearenco A., Berezovsky N. I. 2015. Protein function from its emergence to diversity in contemporary proteins. *Phys. Biol.* 12 045002.
61. González-Segura L, Riveros-Rosas H, JuliaÂn-Sanchez A, Muñoz-Clares RA 2015. Residues that influence coenzyme preference in the aldehyde dehydrogenases. Chem Biol Interact 234: 59-74.
62. Godoy L, Gonzàlez-Duarte R, Albalat R. 2006. S-Nitrosogluthathione reductase activity of amphioxus ADH3: insights into the nitric oxide metabolism. International Journal of Biological Sciences. 2 (3): 117–24
63. Godzik A. 1996. The structural alignment of proteins: is there a unique answer? Protein. Sci. 5:1325-8.
64. Goldstein R. A. 2008. The structure of protein evolution and the evolution of protein structure. Curr. Opin. Struct. Biol. 18:170–177.
65. Greco M, Chiappetta A, Bruno L, & Bitonti M. B. 2012. In *Posidonia oceanica* cadmium induces changes in DNA methylation and chromatin patterning. J. Exp. Bot., 63, 695-709.
66. Gu J, Orr N, Park S.D, Katz L.M, Sulimova G, MacHugh D.E, & Hill E.W. 2009. A genome scan for positive selection in thoroughbred horses. PLoS. One., 4, e5767.
67. Guda C, Guda P, Fahy E, & Subramaniam S. 2004. MITOPRED: a web server for the prediction of mitochondrial proteins. Nucleic Acids Res., 32, W372-W374.

68. Guldan H, Sterner R, Babinger P. 2008. Identification and characterization of a bacterial glycerol-1-phosphate dehydrogenase: Ni(2+)-dependent AraM from Bacillus subtilis. Biochemistry 47: 7376-7384.
69. Guo, H.H; Choe, J; Loeb, L. A. 2004. Protein tolerance to random amino acid change. Proc. Natl. Acad. Sci. U.S.A. 101 (25): 9205–10.
70. Gutheil W.G, Holmquist B, Vallee B.L. 1992. Purification, characterization, and partial sequence of the glutathione-dependent formaldehyde dehydrogenase from Escherichia coli: a class III alcohol dehydrogenase. Biochemistry. 31 (2): 475–81.
71. Guzzo, A.V. 1965. Influence of Amino-Acid Sequence on Protein Structure. Biophys. J. 5 (6): 809-822.
72. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier C.J, & Nakai K. 2007. WoLF PSORT: protein localization predictor. Nucleic Acids Res., 35, W585-W587.
73. Hall B. G. 2011. Phylogenetic trees made easy: A How-to manual. Fourth edition. Sinauer Associates, Sunderland, MA. USA.
74. Hall T. A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids. Symp. Ser. 41, 95-98. 1999.
75. Hellgren M, Strömberg P, Gallego O, Martras S, Farrés J, Persson B, Parés X, Höög J.O 2007. Alcohol dehydrogenase 2 is a major hepatic enzyme for human retinol metabolism. Cellular and Molecular Life Sciences. 64 (4): 498–505.
76. Hernandez-Tobias A, Julian-Sanchez A, Piña E, Riveros-Rosas H. 2011. Natural alcohol exposure: is ethanol the main substrate for alcohol dehydrogenases in animals? Chem. Biol. Interact. 191: 14-25.
77. Hillis, D. M., Bull, J. J. 1993. An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis. Systematic Biology. 42 (2): 182-192.
78. Hiu S.F, Zhu C.X, Yan RT, Chen J.S. 1987. Butanol-Ethanol Dehydrogenase and Butanol-Ethanol-Isopropanol Dehydrogenase: Different Alcohol

Dehydrogenases in Two Strains of Clostridium beijerinckii (Clostridium butylicum). Appl Environ Microbiol 53: 697-703.

79. Holliday G.L, Almonacid D.E, Bartlett G.J. 2007. MACiE (Mechanism, Annotation and Classification in Enzymes): novel tools for searching catalytic mechanisms. Nucleic Acids Res 35:D515–D520.
80. Horton P, Park K.J, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, & Nakai K. 2007. WoLF PSORT: protein localization predictor. Nucleic Acids Res., 35, W585-W587.
81. Hughes A. L. 2007. Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. Heredity. 99:364-73.
82. Huberts D.H., van der Klei I.J. 2010. Moonlighting proteins: an intriguing mode of multitasking. Biochimica et Biophysica Acta. 1803 (4): 520–5.
83. Ingeldew, W. J., and R. K. Poole. 1984. The respiratory chains of *Escherichia coli.* Microbiol. Rev. 48:222-271.
84. Ji-Hyun M., Hyun-Ju L., Suk-Youl P., Jung-Mi S., Mi-Young P., Hye-Mi P., Jiali S., Jeong-Hoh P., Bo Yeon K. and Jeong-Sun K. 2011. Structures of Iron-Dependent Alcohol Dehydrogenase 2 from Zymomonas mobilis ZM4 with and without NAD+ Cofactor. J. Mol. Biol. 407, 413–424.
85. Jeffery C.J. 2003. Moonlighting proteins: old proteins learning new tricks. Trends Genet. 19:415-417
86. Jones D.T., Taylor W.R., Thornton J.M. 1992. A new approach to protein fold recognition. Nature 358 (6381): 86-89.
87. Jörnvall H, Landreh M, Ostberg L.J. 2015. Alcohol dehydrogenase, SDR and MDR structural stages, present update and altered era. Chem Biol Interact. 234: 75-79.
88. Jörnvall H, Harris J.I. 1970. Horse liver alcohol dehydrogenase. On the primary structure of the ethanol-active isoenzyme. European Journal of Biochemistry / FEBS 13(3): 565-576.

89. Jiang H, Blouin C. 2007. Insertions and the emergence of novel protein structure: a structurebased phylogenetic study of insertions. BMC Bioinformatics 8:444.
90. Kardon T, Noel G, Vertommen D, & Schaftingen E.V. 2006. Identification of the gene encoding hydroxyacid-oxoacid transhydrogenase, an enzyme that metabolizes 4-hydroxybutyrate. FEBS Lett., 580, 2347-2350.
91. Kaufman E.E, Nelson T, Fales H.M, & Levin D.M. 1988. Isolation and characterization of a hydroxyacid-oxoacid transhydrogenase from rat kidney mitochondria. J. Biol. Chem., 263, 16872-16879.
92. Kaufman E.E. & Nelson T. 1991. An overview of gamma-hydroxybutyrate catabolism: the role of the cytosolic NADP(+)-dependent oxidoreductase EC 1.1.1.19 and of a mitochondrial hydroxyacid-oxoacid transhydrogenase in the initial, rate-limiting step in this pathway. Neurochem. Res., 16, 965-974.
93. Khersonsky O, Roodveldt C, Tawfik D.S 2006. Enzyme promiscuity: evolutionary and mechanistic aspects. Curr Opin Chem Biol 10:498–508.
94. Kim J.Y, Tillison K.S, Zhou S, Lee J.H, & Smas C.M. 2007. Differentiation-dependent expression of Adhfe1 in adipogenesis. Arch. Biochem. Biophys., 464, 100-111.
95. Kimura M. 1981. Possibility of extensive neutral evolution under stabilizing selection with special reference to nonrandom usage of synonymous codons. Proc. Natl. Acad. Sci. USA 78:5773-7.
96. Kimura M. 1968. Evolutionary rate at the molecular level. Nature. 217:624-6.
97. Kim J.Y, Tillison K.S, Zhou S, Lee J.H, Smas C. M. 2007. Differentiation-dependent expression of Adhfe1 in adipogenesis. Arch. Biochem. Biophys., 464, 100-111.
98. Kovacs B, Stöppler M.C. Alcohol and Nutrition. MedicineNet, Inc. Archived from the original on 23 June 2011. Retrieved 2011-06-07.
99. Koch J. J. 2002. Performance-enhancing: substances and their use among adolescent athletes. Pediatr. Rev., 23, 310-317.

100. Kolaczkowski, B.; Thornton, J. W. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature. 431 (7011): 980-984.

101. Kunin, V., Cases, I., Enright, A. J., De Lorenzo, V., Ouzounis, C. A. 2003. Myriads of protein families, and still counting. Genome Biology. 4 (2): 401.

102. Kumar S. Stecher G., Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Molecular Biology and Evolution. 33 (7): 1887.

103. Lake J. A, Moore J. E. 1998. Trends guide to bioinformatics. Trends. J. Suppl. 1998:22-3.

104. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, & Higgins DG 2007. Clustal W and Clustal X version 2.0. Bioinformatics. 23, 2947-2948.

105. Le S. Q. & Gascuel O. 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol., 25, 1307-1320.

106. Lemey P, Salemi M, Vandamme, A. M. 2009. The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hyphotesis Testing. Cambridge University Press.

107. Letunic, I., Doerks, T., and Bork, P. 2012. SMART 7: recent updates to the protein domain annotation resource. Nucleic. Acids. Res., 40, D302–D305.

108. Li, Wen-Hsiung; Gojobori, Takashi; Nei, Masatoshi 1981. Pseudogenes as a paradigm of neutral evolution. Nature. 292 (5820): 237–9.

109. Lyon R. C, Johnston S. M, Panopoulos A, Alzeer S, McGarvie G, & Ellis E. M. 2009. Enzymes involved in the metabolism of gamma-hydroxybutyrate in SH-SY5Y cells: identification of an iron-dependent alcohol dehydrogenase ADHFe1. Chem. Biol. Interact., 178, 283-287.

110. Madej T, Lanczycki C.J, Zhang D, Thiessen P.A, Geer RC, Marchler-Bauer A, & Bryant S.H. 2014. MMDB and VAST+: tracking structural similarities between macromolecular complexes. Nucleic Acids Res., 42, D297-D303.

111. Maitre M, Klein C, & Mensah-Nyagan A.G. 2016. Mechanisms for the Specific Properties of gamma-Hydroxybutyrate in Brain. Med. Res. Rev., 36, 363-388.

112. Marchler-Bauer, A.; Zheng, C.; Chitsaz, F.; Derbyshire, M. K.; Geer, L. Y.; Geer, R. C.; Gonzales, N. R.; Gwadz, M.; Hurwitz, D. I.; Lanczycki, C. J.; Lu, F.; Lu, S.; Marchler, G. H.; Song, J. S.; Thanki, N.; Yamashita, R. A.; Zhang, D.; Bryant, S. H. 2012. CDD: Conserved domains and protein three-dimensional structure. Nucleic. Acids. Research. 41 (Database issue): D348–D352.

113. Margoliash E. 1963. Primary structure and evolution of cytochrome c. Proc. Natl. Acad. Sci. USA 50:672-9.

114. Marsden R. L., Ranea J.A., Sillero A., 2006. Exploiting protein structure data to explore the evolution of protein function and biological complexity. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 361:425-440.

115. McDonald J.H., Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652-4.

116. Macdonald J.R, Johnson W.C Jr. 2001. Environmental features are important in determining protein secondary structure. Protein Sci. 10(6):1172-7.

117. McGivney B.A, Eivers SS, MacHugh D.E, MacLeod J.N, O'Gorman G.M, Park S.D, Katz L.M, & Hill E.W. 2009. Transcriptional adaptations following exercise in thoroughbred horse skeletal muscle highlights molecular mechanisms that lead to muscle hypertrophy. BMC. Genomics, 10, 638.

118. Miyata, T; Yasunaga, T. 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. Journal of Molecular Evolution. 16 (1): 23–36.

119. Moon JH, Lee HJ, Park SY, Song JM, Park MY, Park HM, Sun J, Park JH, Kim BY, Kim JS. 2011. Structures of iron-dependent alcohol dehydrogenase 2 from *Zymomonas mobilis ZM4* with and without NAD+ cofactor. J Mol Biol 407: 413-424.

120. Montella C, Bellsolell L, Perez-Luque R, Badia 872 J, Baldoma L, Coll M, Aguilar 2005. Crystal structure of an iron-dependent group III dehydrogenase

that interconverts L-lactaldehyde and L-1,2-propanediol in Escherichia coli. J Bacteriol 187: 4957-4966.

121. Morgan, G.J. 1998. Emile Zuckerkandl, Linus Pauling, and the Molecular Evolutionary Clock, 1959-1965. Journal of the History of Biology 31: 155-178.

122. Moss G. P. Recommendations of the Nomenclature Committee. International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes by the Reactions they Catalyse. Retrieved 2006-03-14.

123. Mount D.M. 2004. Bioinformatics: Sequence and Genome Analysis 2nd ed. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY.

124. Nagano N., Nakagawa Z., Arita M., Tsukamoto K., Noguchi T. 2007. "EzCatDB" Nucleic Acids Research, Molecular Biology Database Collection entry number 613.

125. Nelson T, Kaufman E, Kline J, & Sokoloff L. 1981. The extraneural distribution of gamma-hydroxybutyrate. J. Neurochem., 37, 1345-1348.

126. Nei, M; Graur, D. 1984. Extent of protein polymorphism and the neutral mutation theory. Evolutionary Biology. 17: 73-118.

127. Ng, P.C; Henikoff, S. 2006. Predicting the effects of amino acid substitutions on protein function. Annual Review of Genomics and Human Genetics. 7: 61–80.

128. Nimmo, H. G. 1987. The tricarboxylic acid cycle and anaplerotic reactions, p. 156–170. In F. C. Neidhardt, J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology. ASM Press, Washington, D.C.

129. Ojha S, Meng E, Babbitt P, 2007. Evolution of function in the “Two Dinucleotide Binding Domains” flavoproteins. PLoS Comput Biol 3(7):e121

130. Omelchenko MV, Galperin MY, Wolf YI, Koonin EV 2010. Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. Biology Direct. 5: 31.

131. Peng Y, Shi H, Qi X.B, Xiao C.J, Zhong H, Ma R.L, Su B. 2010. The ADH1B Arg47His polymorphism in east Asian populations and expansion of rice domestication in history. BMC Evolutionary Biology. 10: 15.

132. Persson B, Kallberg Y. 2013. Classification and nomenclature of the superfamily of short-chain dehydrogenases/reductases (SDRs). Chem Biol Interact 202: 111-115.

133. Petsalaki EI, Bagos PG, Litou ZI, & Hamodrakas S.J.2006. PredSL: a tool for the N-terminal sequence-based prediction of protein subcellular localization. Genomics Proteomics. Bioinformatics., 4, 48-55.

134. Piatigorsky J, Kantorow M, Gopal-Srivastava R. 1994. Recruitment of enzymes and stress proteins as lens crystallins. EXS 71:241–250.

135. Piatigorsky J, Wistow G.J. 1989. Enzyme/crystallins: gene sharing as an evolutionary strategy. Cell. 57 (2): 197-199.

136. Poole, R. K., and W. J. Ingeldew. 1987. Pathways of electrons to oxygen, p. 170–200. In F. C. Neidhardt, J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology. ASM Press, Washington, D.C.

137. Pundir S, Martin MJ, & O'Donovan C. 2016. UniProt Tools. Curr. Protoc. Bioinformatics., 53, 1.

138. Riveros-Rosas H, Julian-Sanchez A, Villalobos-Molina R, Pardo J.P, Pina E. 2003. Diversity, taxonomy and evolution of medium-chain dehydrogenase/reductase superfamily. Eur. J. Biochem. 270: 3309-3334.

139. Reeves G, Dallman T, Redfern C. 2006. Structural diversity of domain superfamilies in the CATH database. J Mol Biol 360:725–741

140. Reid A.J., Yeats C., Orengo C.A. 2007. Methods of remote homology detection can be combined to increase coverage by 10% in the midnight zone. Bioinformatics 23:2353-2360.

141. Reid M.F, Fewson C.A. 1994. Molecular characterization of microbial alcohol dehydrogenases. Crit Rev Microbiol 20: 13-56.

142. Rigden, Daniel J. From Protein Structure to Function with Bioinformatics. Springer Science. 2009.

143. Rison S.C, Thornton J.M. 2002. Pathway evolution, structurally speaking. Curr. Opin. Struct. Biol. 12:374–382.

144. Rost B. 2002. Enzyme function less conserved than anticipated. J. Mol. Biol. 318:595-608.

145. Ruepp A, Zollner A, Maier D. 2004. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids. Res. 32:5539–5545.

146. Sanghani P.C, Robinson H, Bosron W.F, Hurley T.D. 2002. Human glutathione-dependent formaldehyde dehydrogenase. Structures of apo, binary, and inhibitory ternary complexes. Biochemistry. 41 (35): 10778–86.

147. Saitou N, Nei M. 1986. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, volume 4, issue 4, pp. 406-425.

148. Sengupta S, Ghosh S, Nagaraja V. 2008. Moonlighting function of glutamate racemase from Mycobacterium tuberculosis: racemization and DNA gyrase inhibition are two independent activities of the enzyme. Microbiology. 154 (Pt 9): 2796–803.

149. Shakhnovich B.E., Koonin E.V. 2006 Origins and impact of constraints in evolution of gene families. Genome Res 16:1529-1536.

150. Seibert V, Thiel M, Hinner IS, Schlomann M. 2004. Characterization of a gene cluster encoding the maleylacetate reductase from Ralstonia eutropha 335T, an enzyme recruited for growth with 4-fluorobenzoate. Microbiology 150: 463-472.

151. Sriram G, Martinez J.A, McCabe E.R, Liao J.C, Dipple K.M 2005. Single-gene disorders: what role could moonlighting enzymes play? American Journal of Human Genetics. 76 (6): 911–24.

152. Strickberger, Monroe. 1996. Evolution, 2nd. Ed. Jones & Bartlett.

153. Sultatos L.G, Pastino G.M, Rosenfeld C.A, Flynn E.J. 2004. Incorporation of the genetic control of alcohol dehydrogenase into a physiologically based pharmacokinetic model for ethanol in humans. Toxicological Sciences. 78 (1): 20-31.

154. Tae C.H, Ryu K.J, Kim S.H, Kim H.C, Chun H.K, Min B.H, Chang D.K, Rhee P.L, Kim J.J, Rhee J.C, & Kim Y.H. 2013. Alcohol dehydrogenase, iron containing, 1 promoter hypermethylation associated with colorectal cancer differentiation. BMC. Cancer, 13, 142.

155. Talavera, G., and Castresana, J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Systematic Biology 56, 564-577.

156. Tatusov R.L., Koonin E.V., Lipman D.J. 1997 A genomic perspective on protein families. Science 278:631-637.

157. Taylor W.R., Flores T.P., Orengo C.A. 1994. Multiple protein structure alignment. Protein. Sci. 3(10):1858-70.

158. Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., & Higgins, D.G. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res, 25, 4876-4882.

159. Todd A. E., Orengo C. A., Thornton J. M. 2001 Evolution of function in protein superfamilies, from a structural perspective. J. Mol. Biol. 307:1113–1143.

160. Wallace, I.M.; Blackshields G and Higgins DG. 2005. Multiple sequence alignments. Curr. Opin. Struct. Biol. 15 (3): 261-266.

161. Wang L, Jiang T. 1994. On the complexity of multiple sequence alignment. J. Comput. Biol. 1:337-348.

162. Webb E. C. 1992. Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes. San Diego: Published for the International Union of Biochemistry and Molecular Biology by Academic Press. ISBN 0-12-227164-5.

163. Wheeler T.J., Kececioglu J. D. 2007. Multiple alignment by aligning alignments. Bioinformatics. 23 (13): p. i559-68

164. Whitfield, John B. 1994. ADH and ALDH genotypes in relation to alcohol metabolic rate and sensitivity. Alcohol & Alcoholism supplement 2 pp 61-67

165. Wills C, Kratofil P, Londo D, Martin T. 1981. Characterization of the two alcohol dehydrogenases of *Zymomonas mobilis*. Arch Biochem Biophys 740 210: 775-785.

166. Whelan S, Goldman N. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends Genet. 17(5):262-72. Review.

167. Williamson V.M, Paquin C.E. 1987. Homology of Saccharomyces cerevisiae ADH4 to an iron-activated alcohol dehydrogenase from *Zymomonas mobilis*. Mol Gen Genet 209: 374-381.

168. Van Cauter, E., Plat, L., Scharf, M. B., Leproult, R., Cespedes, S., l'Hermite-Balériaux, M., Copinschi, G. 1997. Simultaneous stimulation of slow-wave sleep and growth hormone secretion by gamma-hydroxybutyrate in normal young Men. Journal of Clinical Investigation. 100 (3): 745–753.

169. Volpi, R., Chiodera, P., Caffarra, P., Scaglioni, A., Saccani, A., Coiro, V. 1997. Different control mechanisms of growth hormone (GH) secretion between γ-amino- and γ-hydroxybutyric acid: neuroendocrine evidence in parkinson's disease. Psychoneuroendocrinology. 22 (7): 531–538.

170. Volpi, R., Chiodera, P., Caffarra, P., Scaglioni, A., Malvezzi, L., Saginario, A., Coiro, V., 2000. Muscarinic cholinergic mediation of the GH response to gamma-hydroxybutyric acid: neuroendocrine evidence in normal and parkinsonian subjects. Psychoneuroendocrinology. 25 (2): 179–85.

171. Vonck J, Arfman N, de Vries GE, Van BJ, van Bruggen EF, Dijkhuizen L. 1991. Electron microscopic analysis and biochemical characterization of a novel methanol dehydrogenase from the thermotolerant Bacillus sp. C1. J Biol Chem 266: 3949-3954.

172. Wallace I M, Blackshields G, Higgins D G. 2005. Multiple sequence alignments. Curr. Opin. Struct. Biol. 15 (3): 261-6. Review.

173. Ward N, Moreno-Hagelsieb G. 2014. Quickly finding orthologs as reciprocal best hits with BLAT, LAST, and UBLAST: how much do we miss? PLoS One 931 9: e101850.

174. Whelan, S. and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum likelihood approach. Molecular Biology and Evolution, 18, 691-699.

175. Williamson V M, Paquin C E. 1987. Homology of Saccharomyces cerevisiae ADH4 to an iron-activated alcohol dehydrogenase from *Zymomonas mobilis*. Mol Gen Genet 209: 374-381. PMID: 2823079

176. Yang, Z. 1996. Maximum-likelihood models for combined analyses of multiple-sequence data. Journal of Molecular Evolution, 42, 587-596.

177. Yang, Z., R. Nielsen, and H. Masami. 1998. Models of amino-acid substitution and applications to mitochondrial protein evolution. Molecular Biology and Evolution, 15, 1600-1611.

178. Yonath, A. 2011. X-ray crystallography at the heart of life science. Current. Opinion in Structural Biology. 21 (5): 622-626.

179. Youngleson J.S, Jones WA, Jones DT, Woods DR. 1989. Molecular analysis and nucleotide sequence of the adh1 gene encoding an NADPH-dependent butanol dehydrogenase in the Gram-positive anaerobe Clostridium acetobutylicum. Gene 78: 355-364.

180. Yuanyuan Z, Yang C, Baishan F. 2004. Cloning and sequence analysis of the dhaT gene of the 1,3-propanediol regulon from *Klebsiella pneumoniae*. Biotechnol Lett 26: 251-255.

181. Zemla A. 2003. LGA - A Method for Finding 3-D Similarities in Protein Structures. Nucleic Acids Research, 31(13):3370-3374.

182. Zhang Y. 2008. Progress and challenges in protein structure prediction. Curr. Opin. Struct. Biol. 18 (3): 342- 348.

183. Zhang Y., Skolnick J. 2005. The protein structure prediction problem could be solved using the current PDB library. Proc. Natl. Acad. Sci. USA 102(4):1029-3.

184. Zhou, T; Gu, W; Wilke, C.O. 2010. Detecting positive and purifying selection at synonymous sites in yeast and worm. Molecular Biology and Evolution. 27 (8): 1912–22.

185. Zou, Dong; Ma, Lina; Yu, Jun; Zhang, Zhang. 2015. Biological databases for human research. Genomics, Proteomics and Bioinformatics. 13 (1): 55-63.

186. Zuckerkandl, E., y Pauling, L. 1962. Molecular disease, evolution, and genetic heterogeneity, pp. 189-225 in Horizons in Biochemistry, edited by M. Kasha and B. Pullman. Academic. Press., New York.

187. Zuckerkandl, E., y Pauling, L. 1965. Evolutionary divergence and convergence in proteins, pp. 97-166 in Evolving Genes and Proteins, edited by V. Bryson and H. J. Vogel. Academic. Press., New York.