



Trabajos Prácticos Bioinformática 2021

Editores:

Dr. Mauricio J. Lozano

Dra. M. Leticia Ferrelli

Prof. Dr. Gustavo Parisi

Trabajos prácticos Bioinformática 2021 editado por María Leticia Ferrelli ; Mauricio Lozano; Gustavo Parisi. - 1a ed. - La Plata : Universidad Nacional de La Plata. Facultad de Ciencias Exactas, 2022.

Libro digital, PDF

Archivo Digital: descarga y online
ISBN 978-950-34-2180-2

CDD 572.636

Trabajos Prácticos Bioinformática 2021

Editores:

Dr. Mauricio J. Lozano

Dra. M. Leticia Ferrelli

Prof. Dr. Gustavo Parisi

Presentación

La Bioinformática es un área interdisciplinaria que se ocupa del análisis computacional de los sistemas biológicos, siendo una de sus ramas la aplicación de este tipo de análisis a los sistemas moleculares. Si bien una parte de la Bioinformática se ocupa del desarrollo de nuevas metodologías, en la actualidad contamos con un gran conjunto de herramientas computacionales que permiten sistematizar, extraer y analizar la información biológica contenida en secuencias moleculares tanto de ácidos nucleicos como de proteínas. Estas herramientas permiten entre otras cosas, predecir la estructura de proteínas, diseñar ligandos específicos como inhibidores o antibióticos, diseñar racionalmente proteínas, identificar sitios funcionales y predecir la función biológica. Así, la Bioinformática es un campo estrechamente relacionado con la biotecnología, la bioquímica, la biología molecular, la farmacología, y consecuentemente tiene incidencia en distintas áreas como por ejemplo la salud y el agro, tanto en el ámbito académico como industrial.

La incorporación de estas herramientas acompañada del marco conceptual adecuado para su utilización y su articulación con resultados experimentales, son los principales objetivos de la asignatura optativa -y de postgrado- Bioinformática, perteneciente a la Licenciatura en Biotecnología y Biología Molecular dictada por el Área Biotecnología y Biología Molecular de la Facultad de Ciencias Exactas, Universidad Nacional de La Plata. Los docentes a cargo de la materia son el Profesor Gustavo Parisi y los JTPs Mauricio Lozano y María Leticia Ferrelli. El programa de la materia incluye los siguientes temas orientados principalmente al estudio de proteínas: estructura de proteínas, evolución biológica, estudios de similitud secuencial, utilización de bases de datos biológicas, estimación de la estructura de proteínas, estudios basados similitud estructural, modelado molecular, inferencia filogenética, e integración estructura-evolución (predicción de la función biológica).

En el contexto de esta asignatura, y con el objetivo de que los estudiantes adquirieran una mayor experiencia práctica en las diferentes temáticas estudiadas, se realizó un trabajo práctico integrador desarrollado a lo largo de 12 clases, durante las cuales se profundizó en la caracterización de una proteína que fue elegida por los estudiantes bajo la supervisión de la cátedra. Como resultado del análisis bioinformático realizado se exigió a los estudiantes la presentación de un trabajo escrito individual, con el objetivo de fijar los conocimientos adquiridos, e introducir a los estudiantes en la escritura científica.

Previamente a la entrega final se realizaron dos etapas de evaluación y corrección que fueron utilizadas como evaluaciones parciales de los conocimientos prácticos adquiridos. En estas instancias se evaluaron y corrigieron los métodos utilizados y la interpretación de los resultados obtenidos. Los contenidos incluidos en las etapas 1 y 2 fueron los siguientes:

Etapas 1: Análisis secuencial

- Búsqueda de información de la proteína en diferentes bases de datos. A partir de esta información, y una breve búsqueda bibliográfica, deberá redactarse la introducción.
- Búsqueda de homólogos cercanos y remotos utilizando diferentes métodos (Secuencia-secuencia, profile-secuencia, HMM). A partir de los resultados deberá presentar un análisis del número de homólogos y su distribución taxonómica.
- Predicción de estructura secundaria, segmentos transmembrana, regiones desordenadas y dominios.

- Selección de un conjunto de secuencias (según el objetivo de análisis planteado) con las cuales deberá realizarse un alineamiento múltiple.

Etapa 2: Análisis estructural y filogenético

- Obtención de un modelo molecular para la proteína. Esta etapa deberá detallar el proceso utilizado incluyendo: Asignación de plegamiento, alineamiento, modelado, visualización. Comparación con el template y evaluación de la calidad del modelo.
- Obtener la clasificación estructural de la proteína.
- Obtención de un árbol filogenético por el método de máxima verosimilitud con soporte de las ramas por bootstrap. En esta etapa se deberá detallar el proceso utilizado, incluyendo la evaluación del alineamiento, selección del modelo evolutivo, parámetros utilizados para la generación del árbol.
- Realizar un análisis funcional utilizando tanto datos estructurales como secuenciales.
- Se redactará la conclusión a partir de la información de la primera y segunda parte.

Finalmente se corrigió y unificó el formato al de una comunicación científica estructurada en las siguientes secciones: Introducción, Objetivos, Métodos y resultados, Conclusiones, Referencias.

En el presente libro se recopilan los trabajos realizados por estudiantes de la cursada 2021. Cada trabajo fue realizado a criterio de cada estudiante y con la posibilidad de contestar preguntas específicas que surgieran en el análisis de dicha proteína. Las herramientas vistas en el curso son numerosas y cada autor/a del trabajo utilizó las que consideraba convenientes. El conjunto de proteínas estudiadas en este libro corresponden a diversos organismos, incluyendo especies bacterianas, arqueas, eucariotas y virus:

ID Uniprot	Proteína	Organismo	Super-reino
A0A6V8D8A1	Thymidylate synthase	<i>Candidatus Poseidoniales archaeon</i>	Archaea
Q89KW9	Bll4781 protein	<i>Bradyrhizobium diazoefficiens</i>	Bacteria
B5F500	Phosphoglycerol transferase I	<i>Salmonella agona</i>	Bacteria
Q6W3M3	ATP citrate synthase	<i>Alvinella pompejana epibiont</i>	Bacteria
Q8TCU5	Glutamate receptor ionotropic, NMDA 3A	<i>Homo sapiens</i>	Eukaryota
Q2TA06	Aurora kinase A	<i>Bos taurus</i>	Eukaryota
Q96D96	Voltage-gated hydrogen channel 1	<i>Homo sapiens</i>	Eukaryota
Q91V45	KISS-1 receptor	<i>Mus musculus</i>	Eukaryota
-----	metacaspasa-4 (MCA4)	<i>Nicotiana benthamiana</i>	Eukaryota

Índice

Presentación	4
Índice	8
Análisis secuencial de la proteína fosfoglicerol transferasa I de <i>Salmonella agona</i> (cepa SL 483). <i>Déborah Colman</i>	12
RESUMEN	12
INTRODUCCIÓN	12
MÉTODOS	13
RESULTADOS	14
CONCLUSIONES Y DISCUSIÓN	18
BIBLIOGRAFÍA	19
Análisis bioinformático de una cisteín proteasa presente en plantas de <i>Nicotiana benthamiana</i>, la proteína Metacaspasa 4. <i>Ana Marchesini</i>	21
RESUMEN	21
INTRODUCCIÓN	21
MÉTODOS Y RESULTADOS	22
Análisis de secuencia	22
Análisis de homología	24
Modelado de la proteína: estructura terciaria	25
Análisis taxonómico	26
Predicción de la función	27
CONCLUSIONES Y DISCUSIÓN	29
BIBLIOGRAFÍA	29
Estudio bioinformático de la proteína Aurora kinasa A de <i>Bos Taurus</i>. <i>Josefina Ormaechea</i>	31
RESUMEN	31
INTRODUCCIÓN	31
MÉTODOS Y RESULTADOS	32
PRIMERA PARTE: Análisis secuencial	32
SEGUNDA PARTE: Análisis estructural y filogenético	35
CONCLUSIONES Y DISCUSIÓN	40
BIBLIOGRAFÍA	40
Modelado molecular y predicción funcional de la proteína bll4781 de <i>Bradyrhizobium diazoefficiens</i> (USDA 110). <i>Damián Brignoli</i>	42
RESUMEN	42
INTRODUCCIÓN	42
MÉTODOS Y RESULTADOS	43
CONCLUSIONES Y DISCUSIÓN	47
BIBLIOGRAFÍA	47

Implicancias del receptor ionotrópico NMDA subunidad 3A en la esquizofrenia. Iris Quimey López	49
RESUMEN	49
INTRODUCCIÓN	49
Hipótesis sobre la esquizofrenia	50
MÉTODOS Y RESULTADOS	51
Búsqueda de secuencias homólogas	51
Predicción de estructura secundaria	52
Regiones de baja complejidad	53
Predicción de dominios globulares	54
Predicción de segmentos transmembrana	54
Predicción de regiones coiled-coil	55
Predicción de regiones desordenadas	56
Búsqueda de regiones repetitivas	57
Búsqueda de dominios conservados y motivos secuenciales	57
Asignación de plegamiento	60
Modelado por homología utilizando Modeller	64
Estimación filogenética	65
Predicción de función	66
CONCLUSIONES Y DISCUSIÓN	68
BIBLIOGRAFÍA	70
Estudios bioquímicos de la ATP citrato liasa de bacterias simbiotes de <i>Alvinella pompejana</i> (Anélida: Poliqueta). Ignacio Pavía	73
RESUMEN	73
INTRODUCCIÓN	73
MÉTODOS Y RESULTADOS	74
Predicción de elementos estructurales	74
Búsqueda de homólogos	74
Alineamiento múltiple	74
Modelado por homología	74
Inferencia filogenética	77
CONCLUSIONES Y DISCUSIÓN	78
BIBLIOGRAFÍA	78
Recorrido por Hv1: un canal selectivo de protones	79
RESUMEN	79
INTRODUCCIÓN	79
MÉTODOS Y RESULTADOS	80
CONCLUSIONES Y DISCUSIÓN	86
BIBLIOGRAFÍA	86
Análisis bioinformático de la enzima Timidilato Sintasa de <i>Candidatus Poseidoniales archaeon</i>. Bernarda Pschunder	87
RESUMEN	87
INTRODUCCIÓN	87

MÉTODOS Y RESULTADOS	87
Búsqueda de homólogos cercanos y remotos utilizando diferentes metodologías	87
Predicción de estructura secundaria, segmentos transmembrana, regiones desordenadas y dominios	89
Clasificación estructural de la proteína	91
Obtención de un modelo molecular para la proteína	91
Selección y alineamiento múltiple de secuencias	93
Obtención de un árbol filogenético	94
Análisis funcional de la Timidilato Sintasa	95
CONCLUSIONES Y DISCUSIÓN	97
BIBLIOGRAFÍA	98
Caracterización secuencial, estructural y evolutiva del receptor de Kisspeptina, GPR-54, en <i>Mus musculus</i> mediante herramientas bioinformáticas. <i>Alejandro Raúl Schmidt</i>	99
RESUMEN	99
INTRODUCCIÓN	100
MÉTODOS Y RESULTADOS	101
CONCLUSIONES Y DISCUSIÓN	110
BIBLIOGRAFÍA	111
Conclusiones	113
Apendice	115
A1 - Material suplementario: Análisis secuencial de la proteína fosfoglicerol transferasa I de <i>Salmonella agona</i> (cepa SL 483)	116
A2 - Material suplementario: Análisis bioinformático de una cisteín proteasa presente en plantas de <i>Nicotiana benthamiana</i> , la proteína Metacaspasa 4.	119
A3 - Material suplementario: Modelado molecular y predicción funcional de la proteína bll4781 de <i>Bradyrhizobium diazoefficiens</i> (USDA 110)	126
A4 - Material suplementario: Implicancias del receptor ionotrópico NMDA subunidad 3A en la esquizofrenia	129
A5 - Material suplementario: Recorrido por Hv1: un canal selectivo de protones	138
A6 - Material suplementario: Análisis bioinformático de la enzima Timidilato Sintasa de <i>Candidatus Poseidoniales archaeon</i>	140

Análisis secuencial de la proteína fosfoglicerol transferasa I de *Salmonella agona* (cepa SL 483)

Déborah Colman

Cátedra de Bioinformática, Área de Biotecnología y Biología Molecular, Departamento de Ciencias Biológicas, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, La Plata, Argentina.

RESUMEN

La enzima fosfoglicerol transferasa I (gen *mdoB*) de la enterobacteria *Salmonella agona* (cepa SL483) fue caracterizada a partir de una secuencia de aminoácidos mediante distintas herramientas bioinformáticas. Datos bibliográficos de otros microorganismos asociados filogenéticamente determinaron que la función biológica podría participar del traspaso de azúcares dentro del sistema fosfotransferasa. Nuestros resultados indicaron que la proteína está compuesta por 763 aminoácidos con peso molecular de 85084 Da. Se determinó que la secuencia aminoacídica en estudio tiene, al menos, un plegamiento secundario conformado por un péptido señal, 4 regiones transmembrana y un dominio Sulfatasa. El plegamiento de este dominio conforman 4 alfas-hélices hacia el exterior y 4 hojas beta en el interior, dispuestas en un arreglo globular. Se buscaron proteínas homólogas cercanas y remotas con el fin de investigar las variaciones evolutivas. Los géneros taxonómicos más representados fueron: *Salmonella* sp., *Escherichia* sp., *Shigella* sp. y *Citrobacter* sp.. Los resultados alcanzados permitieron predecir la función proteica.

PALABRAS CLAVE: proteína - dominio - estructura secundaria - plegamiento.

INTRODUCCIÓN

Las enterobacterias tienen membrana fosfolipídica interna como externa, formando así un espacio periplásmico contenedor de la pared celular formada por peptidoglicano. La membrana interna o membrana citoplasmática es impermeable a las moléculas polares, regula el paso de nutrientes, metabolitos y macromoléculas, además mantiene la fuerza motriz protónica que es fundamental en el metabolismo energético bacteriano; mientras que el espacio periplásmico contiene una gran concentración de proteínas y peptidoglucano (Quirós Cárdenas).

La proteína fosfoglicerol transferasa I es una enzima glicerotransferasa de la membrana citoplasmática de bacterias Gram negativas que forma parte del sistema fosfotransferasa responsable de la incorporación de azúcares. La función biológica está relacionada con el metabolismo de glicerolípidos, principalmente en la transferencia de residuos fosfoglicerol desde el fosfatidilglicerol hasta oligosacáridos derivados de membrana (membrane-derived oligosaccharides, MDO), constituyentes del espacio periplásmico (Jackson et al.). Estudios previos reportaron que el sitio activo de esta proteína se ubica sobre la región periplásmica de la membrana interna y su actividad está estrictamente regulada por la osmolaridad (Bohin et al.).

Las primeras caracterizaciones de esta proteína se hicieron usando como modelo a *Escherichia coli*, pero también se han reportado en otras bacterias Gram negativas, tal como *Salmonella* sp. (Schulman, et al.; [Fricke](#), et al.). El objetivo de este análisis es caracterizar la proteína a partir de su secuencia de aminoácidos y comparar los cambios en las secuencias de organismos relacionados evolutivamente utilizando herramientas bioinformáticas.

MÉTODOS

Identificación de la proteína. Se utilizó la plataforma de identificación de proteínas Uniprot, la secuencia query estaba anotada con el código B5F500. Otra alternativa fue el programa de búsqueda por homología Blastp contra la base de datos no redundantes. Los parámetros configurados fueron los establecidos por default (Max target sequences 5000, e-value threshold 0.05, Word size 6, Matrix BLOSUM62).

Búsqueda de homólogos cercanos y remotos. La búsqueda de homólogos cercanos se hizo con Blastp (NCBI) y Blast de UniProt contra la base de datos Swiss-Prot. La búsqueda de homólogos remotos se realizó con el servidor online HMMER, utilizándose el programa jackhmmmer contra la base de datos UniProt - Reference Proteomes. Rango de e-values configurado: 0.00001 – 0.03. Se hicieron 2 iteraciones.

Caracterización de secuencia aminoacídica. Predicción de estructura secundaria. La base de datos Pfam se utilizó para caracterizar a la proteína y en base a la información brindada por los links (InterPro, Phobius) de otros servidores se recopilaron más datos. Para estimar la presencia de desorden se usó Iupred2A seteado en IUPred Structured domains. Para detectar regiones flexibles, Dynamine fue la opción elegida. El servidor JPred se utilizó para caracterizar la estructura secundaria basada en la composición de aminoácidos.

Alineamiento múltiple. Para seleccionar las secuencias del alineamiento múltiple, se tomaron los hits provenientes del análisis realizado con el programa jackhmmmer para búsquedas de homólogos remotos. Se revisaron los alineamientos y parámetros y luego se elaboró un archivo multifasta conteniendo 31 secuencias hits junto con la secuencia query. El alineamiento múltiple (Multiple sequence alignment, MSA) se construyó con el programa Toffee.

Asignación de plegamiento. A partir de la secuencia aminoacídica se utilizó la herramienta HHpred para predecir la estructura proteica. Se utilizaron los parámetros seteados por default.

Búsqueda del template para modelar. Se realizó la búsqueda de proteínas homólogas con estructura conocida para modelar la secuencia query. Se usó el programa de búsqueda remota HHpred con los parámetros seteados por default para la asignación de plegamiento.

La revisión del alineamiento de la secuencia query y de la secuencia template se hizo con el programa Notepad++, de esta manera los missing residues y gaps fueron corregidos manualmente. El programa de modelado por homología elegido fue el Modeller versión 10.1. Se obtuvieron 10 modelos, y luego de revisar los potenciales energéticos (Dope score y moldpdf) se seleccionó aquel con menores puntajes. La evaluación del modelo seleccionado se realizó con el programa ProSA, el cual analiza la energía global y por posición. También se optó por Dope, integrado en el programa Modeller.

Alineamiento estructural. A partir de la estructura generada, el programa PyMol fue el señalado para alinear las estructuras proteicas query-template. Se calculó el RMSD.

Estimación filogenética. A partir del alineamiento múltiple realizado previamente, se procedió a analizar el modelo de evolución más ajustado a las secuencias aminoacídicas. Para ello se utilizó el subprograma MODELTEST incorporado en el paquete HyPhy. Se construyó un árbol filogenético por Neighbor Joining para comparar entre modelos.

Con el programa PHYML se estimó la filogenia mediante el método de máxima verosimilitud (maximum likelihood, ML), utilizando el modelo evolutivo designado previamente. El análisis de ML se determinó con un soporte de las ramas por Bootstrap = 100, se consideró la elaboración del modelo con el parámetro gamma distribution. Se utilizó una secuencia outgroup, cuya selección se determinó mediante búsqueda

por homología en el programa blastp contra la base de datos que incluía únicamente el taxón Archaea (MBC8501365). El nuevo alineamiento múltiple se repitió como fue detallado anteriormente.

Predicción de función. Para predecir sitios funcionales de la proteína en estudio se utilizaron los servidores ConSurf y Evolutionary Trace.

RESULTADOS

Identificación de la proteína. La identificación de la proteína en Uniprot (compuesta por 763 aminoácidos, peso molecular 85084 Da) arrojó que se trata de la enzima fosfoglicerol transferasa I, también referenciada como glicerofosfotransferasa fosfatidilglicerol-oligosacárido de membrana, codificada en el gen *mdoB* de la enterobacteria *Salmonella agona* (cepa SL483). Su función biológica es transferir residuos de fosfoglicerol desde fosfatidilglicerol a la cadena de carbonos de glucanos unida a membrana. La actividad catalítica es (EC=[2.7.8.20](#)): Fosfatidilglicerol + oligosacárido derivado de membrana D-glucosa \rightleftharpoons 1,2-diacil-sn-glicerol + oligosacárido derivado de membrana 6-(glicerofosfo)-D-glucosa. La ruta metabólica asociada es la biosíntesis de glucano periplásmico osmoregulado (osmoregulated periplasmic glucan, OPG).

Se trata de una proteína globular conformada por un péptido señal, cuatro regiones transmembrana y un dominio identificado como Sulfatasa (Figura 1), perteneciente a la Superfamilia de las Fosfatasas alcalinas.



Figura 1. Características estructurales (Pfam) de la proteína *mdoB* de *S. agona* (Uniprot ID: B5F500).

Por su parte, la búsqueda por homología en el programa Blastp contra la base de datos no redundante informó que la secuencia aminoacídica es la proteína glicerofosfotransferasa fosfatidilglicerol-oligosacárido de membrana (MULTISPECIES: phosphatidylglycerol--membrane-oligosaccharide glycerophosphotransferase [*Salmonella*]), en concordancia con lo detectado en Uniprot. El identificador de secuencia es WP_001292705.1. Los valores del match fueron: Score 1579 bits, e-value 0.0, Porcentaje de identidad (%ID): 100%, Porcentaje de similitud: 100% y Gaps 0%. El número de hits obtenidos en las condiciones analizadas fue de 4985, cuyo porcentaje de identidad abarcó el rango de 82.57% hasta 100% con un e-value de 0.0 para todos los hits. La taxonomía presentó más del 95% organismos clasificados dentro de la familia Enterobacteriaceae (clase gamma-Proteobacteria). Los géneros taxonómicos más representados fueron: *Salmonella* sp., *Escherichia* sp. *Shigella* sp. y *Citrobacter* sp..

Búsqueda de homólogos cercanos y remotos. En función de la poca diversidad taxonómica conseguida con Blastp, se optó por usar blast contra la base de datos Swiss-Prot, obteniéndose un total de 48 hits; de los cuales, 37 matches comprendieron un rango de %ID entre 38.3% hasta 100%. Los 11 hits restantes fueron desestimados después de analizar los alineamientos y parámetros resultantes (%ID menor al 30%, e-value $5.8e^0 - 8.3e^{-11}$, scores 77-168).

El grupo con 37 hits se dividió en 2 subgrupos de acuerdo con el %ID. Uno de los subgrupos contenía hits con un %ID del 40% aproximadamente, y e-value en el rango de $8.2e^{-82} - 1.1e^{-83}$, siendo *Xanthomonas* sp. el único género taxonómico; mientras que el otro subgrupo estuvo comprendido entre el 90 - 100% de identidad, e-value 0.0 y cuyos taxones correspondieron a los taxones *Salmonella* sp., *Escherichia* sp. *Shigella* sp. y *Citrobacter* sp.,.

La búsqueda de homólogos remotos en jackhammer resultó en 2126 hits totales, de los cuales, 63 hits únicamente dieron parámetros considerables y, con valores de %ID no estrictamente correctos o esperables. Es decir, se seleccionaron hits con e-values en el rango de $2.9e^{-272}$ hasta 0.0, con %ID cercano al 50%, % de similitud mayor al 70%, y Bit score superior a 900. Los taxones correspondieron al phylum Proteobacteria [familia Enterobacteriaceae. Géneros taxonómicos: *Salmonella* sp., *Escherichia* sp., *Citrobacter* sp., *Shigella* sp., *Kluyvera* sp., *Klebsiella* sp., *Enterobacter* sp., *Superficieibacter* sp., *Raoultella* sp., *Erwinia* sp., *Pantoea* sp., *Serratia* sp.]. Dados los porcentajes de identidad y similitud obtenidos con esta estrategia se concluye que los hits encontrados aún corresponden a homólogos cercanos.

Caracterización de regiones secuenciales y predicción de estructura secundaria. A partir de lo obtenido con Pfam, la fosfoglicerol transferasa I se conforma por un péptido señal (sitio 1-18); 4 regiones transmembrana (sitios 28-47; 59-74; 80-98; 110-128), un dominio sulfatasa (sitio 163-446). El dominio Sulfatasa (código de acceso PF00884.23) pertenece al clan CL0088. A partir de las referencias cruzadas de Pfam, se pudo acceder a la base de datos InterPro que confirmó que la proteína en estudio tiene un dominio Sulfatasa y pertenece a la Superfamilia de las fosfatasa alcalinas. El análisis con Phobius confirmó la presencia de las regiones transmembrana (ver Anexo).

El programa IUPred2A reportó que la enzima de interés tiene estructura globular. Se realizó una búsqueda de motivos secuenciales en Prosite, pero no se obtuvieron resultados. Se evaluó el desorden de esta proteína mediante la herramienta bioinformática Dynamine, estimando una estructura poco flexible, lo cual es propio de las proteínas globulares (ver Anexo).

La predicción de la estructura secundaria se realizó con el programa JPred, cuyo resultado estimó que hacia el extremo N-terminal de la secuencia de aminoácidos se forma un arreglo tipo alfa hélice hasta la posición 140 aproximadamente, luego le siguen cortos segmentos de arreglos alfa hélice y hojas betas: y desde la posición 600 aproximadamente hasta el extremo C-terminal, el programa predijo que la secuencia se ordenó en hojas beta (Figura 2).



Figura 2. Resultado de la predicción de la estructura secundaria con el programa JPred. Los segmentos rojos corresponden a arreglos alfa hélices y las flechas verdes corresponden a arreglos hoja beta.

Comparación con secuencias homólogas. Alineamiento múltiple. Para realizar el MSA se seleccionaron 31 hits de los determinados previamente en la búsqueda de homólogos. Las secuencias para alinear se eligieron según los parámetros de confiabilidad (e-value, %ID, %similitud, Score), y la diversidad taxonómica. En el MSA se observó que la única región transmembrana de la proteína de interés que se conservó es la posicionada entre 59-74 (Figura 3.A). Respecto a la región del dominio Sulfatasa se encontró altamente conservado para los organismos elegidos (Figura 3.B).

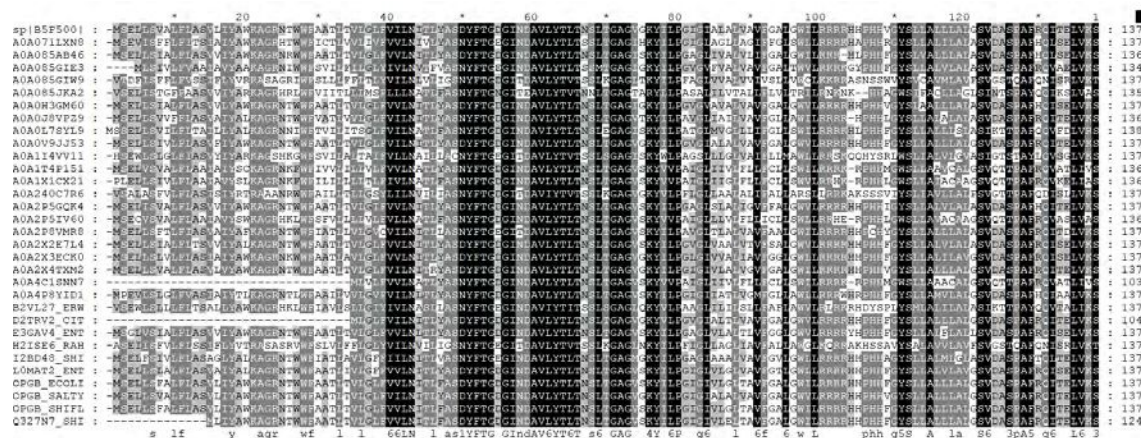


Figura 3.A. Alineamiento múltiple. La imagen corresponde a la región transmembrana conservada (59-74). Las referencias de los organismos elegidos se enlistan en el Anexo.



Figura 3.B. Alineamiento múltiple. La imagen corresponde al dominio Sulfatasa conservado. Las referencias de los organismos elegidos se enlistan en el Anexo.

Análisis estructural. El análisis de la proteína fosfoglicerol transferasa I prosiguió con la estimación de su estructura tridimensional.

Para realizar la asignación de plegamiento se utilizó la herramienta HHpred después de probar con varias opciones bioinformáticas, tal como Phyre2. El alineamiento con HHpred determinó que el mejor hit fue la proteína 3LXQ de PDB (correspondiente a la entrada Q87NY2 en Uniprot), identificada como [Vibrio parahaemolyticus](#) serotype O3:K6 (Gammaproteobacteria). Contiene un dominio Sulfatasa con unión al ion manganeso. Los valores fueron 25 % ID, e-value $3e^{-29}$, score 278.32, % de similitud 28.9. El template alineó desde la posición 159 hasta la posición 460 de la secuencia proteica query, coincidiendo con la región del dominio Sulfatasa. La resolución de la cristalografía fue de 1.95 Å. Esta estructura proteica dio mejor valoración para usarse como molde para la proteína en estudio. La cadena A de la secuencia molde correspondió al alineamiento con la secuencia en estudio.

Para iniciar el modelado molecular para la proteína se utilizó el template propuesto por HHpred en el análisis de fold assignment. Aun considerando que el modelo a generar no sería el óptimo para determinaciones más precisas, se eligió este template.

Cabe mencionar que se probaron otros métodos de búsquedas por homología tales como Blastp contra PDB, Psiblast y FFas03, pero ninguna de estas herramientas arrojó mejores hits y/o valores que la herramienta mencionada más arriba.

El programa de modelado fue seteado para determinar 10 modelos, de los cuales se eligió uno a partir de los potenciales estadísticos intrínsecos del programa. Se analizó el modelo mediante dos parámetros energéticos: global (z-score: -6.35, Prosa) y local (Dope score) (Figura 4).

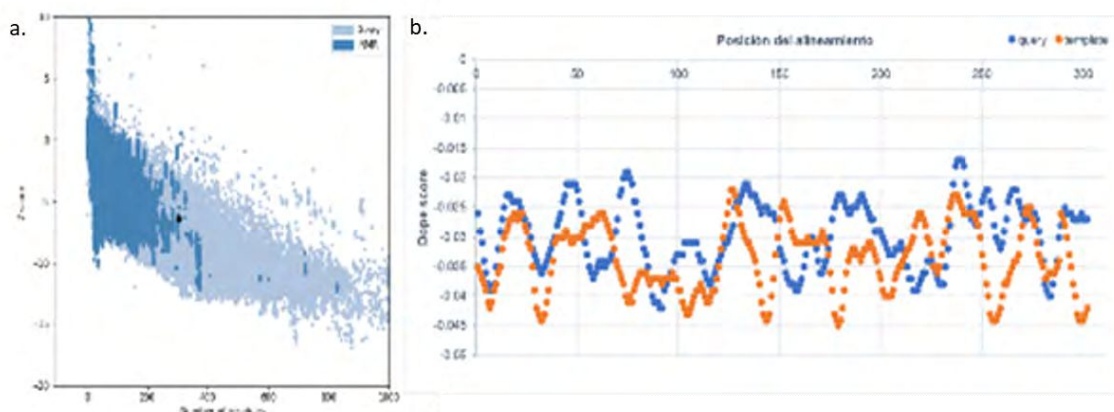


Figura 4. Análisis energético del modelo tridimensional propuesto para la proteína en estudio. a. Validación energética global por el programa ProSA. El punto negro refiere a la proteína de interés. b. Perfil energético por sitio superpuesto de las secuencias query (azul) y template (naranja).

El modelo demostró que la proteína query tiene al menos 4 segmentos alfa hélices hacia el exterior de la estructura, y 4 hojas beta en el interior; también se observan loops que no están alineados (Figura 5 a, b y c). En términos generales, tiene forma globular. A partir del alineamiento estructural se observó que 450 aminoácidos del template fueron alineados contra 302 aminoácidos de la secuencia query, estimando un score match align de 236.5. El cálculo de RMSD arrojó un valor de 0.137, según el programa PyMol.

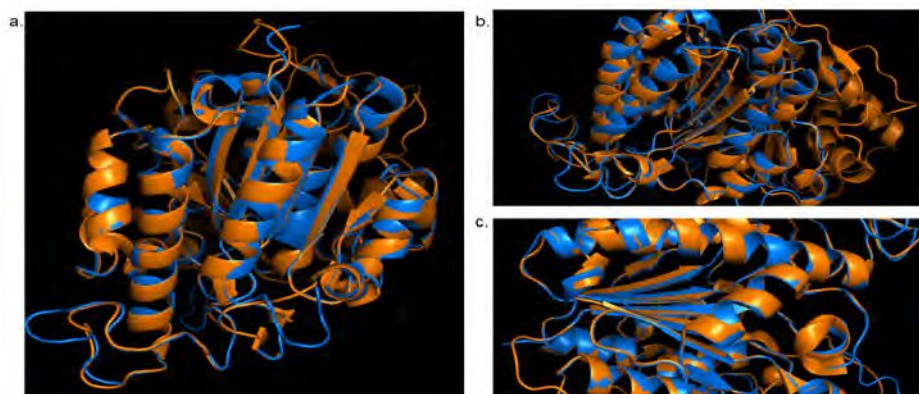


Figura 5. Alineamiento estructural. Estructura tridimensional de proteína query (azul) superpuesta con la estructura de la proteína molde (naranja). a. Se observa el plegamiento alfa hélice hacia la superficie de la estructura. En b. y c. se muestran los plegamientos hojas beta en el interior de la cavidad proteica.

Estimación filogenética. Utilizando las secuencias del MSA descrito anteriormente se realizó una estimación filogenética. El resultado de la prueba arrojó que el modelo de evolución más adecuado para las secuencias a estudiar fue WAG + F. El outgroup elegido no fue el correcto pues marca mucha distancia evolutiva del resto de las secuencias. La secuencia query fue determinada en un nodo con buen soporte (arrojó un bootstrap de 100), por lo que la información que contiene el alineamiento fue suficientemente robusta para confirmar la relación evolutiva en ese nodo (Figura 6).

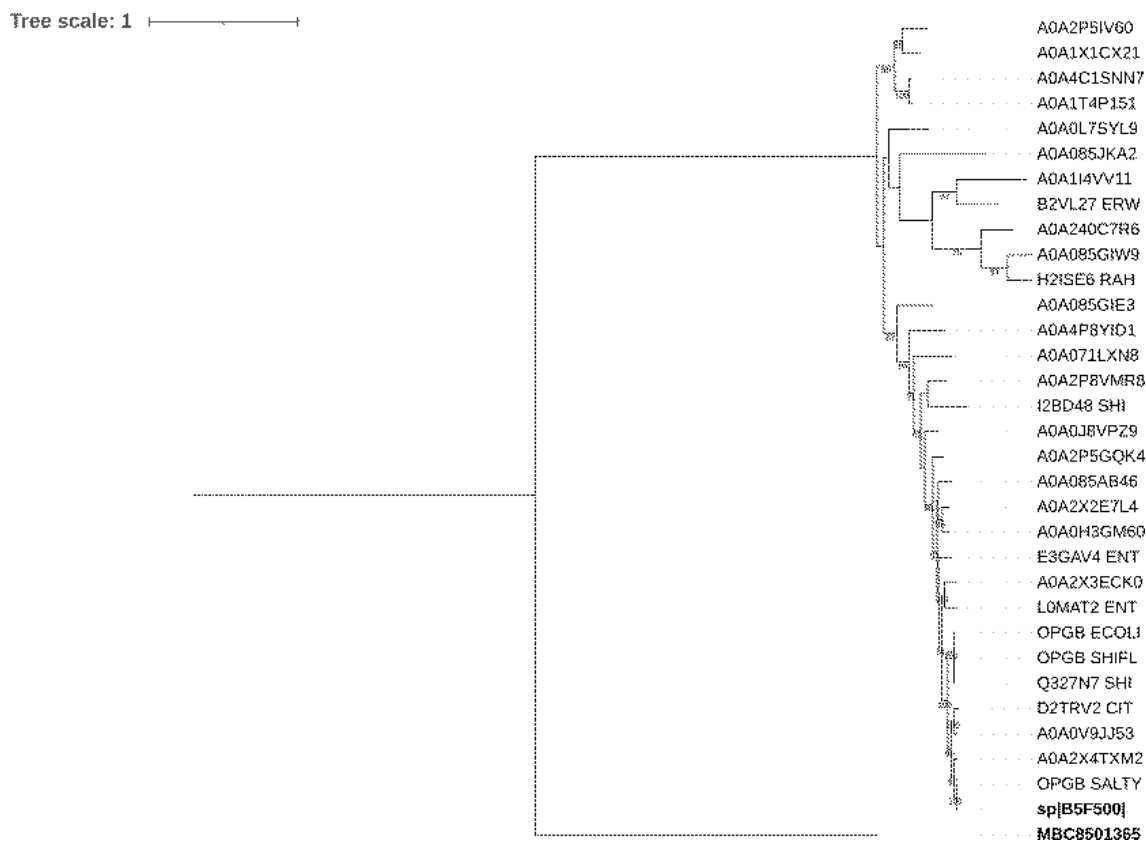


Figura 6: Árbol filogenético construido por el método de máxima verosimilitud. Bootstrap=100.Outgroup: secuencia aminoacídica de Archaea (MBC8501365). Bootstraps mostrados corresponden a valores > 50.

Predicción de función. El análisis de ConSurf indicó que la estructura de la proteína tiene en sus hélices alfa regiones más conservadas, que podrían estar asociadas a sitios importantes para la función, cuya disposición se indica hacia el interior de la proteína, mientras que las regiones menos conservadas se posicionan en el exterior de la estructura tridimensional tal como era de esperarse (Figura 7). Respecto de los scores más altos de conservación, estos están en el centro de la proteína y se sitúan las hojas beta y hacia el exterior se ubican los arreglos alfa hélices.

CONCLUSIONES Y DISCUSIÓN

A partir de estas determinaciones podríamos sugerir que la proteína fosfoglicerol transferasa I tiene estructura globular, con al menos 4 arreglos alfa hélices hacia el exterior, dejando hacia el core al menos 4 arreglos hojas beta. Los residuos que componen cada una de estas estructuras secundarias se dispusieron de acuerdo a su conservación evolutiva. Aquellos más conservados se ubican hacia el interior de la proteína globular, mientras que los residuos menos conservados, se encuentran hacia el exterior.



Figura 7. Estructura de proteína fosfoglicerol transferasa I. La escala de colores señala la conservación de los residuos (rojo) hasta los residuos menos conservados (violeta).

Respecto a la estructura no se pudo evaluar los loops, por lo que será necesario re evaluar el template para intentar estimar algo al respecto. En términos generales, la estructura terciaria lograda podría considerarse buena, dado el bajo porcentaje de identidad del template, aunque insuficiente para lograr un acabado modelado de la proteína. Este molde se seleccionó entre los pocos hits arrojados por las distintas bases de datos estructurales analizadas.

Es importante mencionar que el análisis de la estructura terciaria se realizó a partir de la cadena A del template cuyo alineamiento fue contra la región del dominio Sulfatasa, situado en las posiciones 159-460 de la secuencia query, dicha ubicación coincidió con lo informado por Pfam (dominio Sulfatasa: 163-446). Los arreglos descritos en la estructura secundaria, más precisamente en la zona transmembrana en el N-terminal no pudieron visualizarse en el análisis tridimensional dado que no se obtuvo un template que cubra esa región. Sería de mera importancia conseguir en un próximo estudio un molde que supere las limitaciones expuestas en este trabajo.

Como ya se ha mencionado, la proteína es globular, lo cual es típica de proteínas de membrana. Respecto al alineamiento múltiple para el análisis filogenético, se puede decir que las secuencias elegidas pertenecieron al grupo taxonómico enterobacterias, por lo que el outgroup elegido fue incorrecto, generando importante distancia evolutiva. Asimismo, no hubo una notoria separación de ramas, esto pudo deberse a que las secuencias seleccionadas correspondieron al mismo taxón.

BIBLIOGRAFÍA

Bohin Jean-Pierre & Kennedy E. P. (1984). Regulation of the Synthesis of Membrane-derived Oligosaccharides in *Escherichia coli*. *J. Biol. Chem.*, 13(259), 8388-8393.

Fricke W., [Mammel M. K.](#), [McDermott P. F.](#), [Tartera C.](#), [White D. G.](#), [Leclerc, J. E.](#), [Ravel J.](#), [Cebula T. A.](#) (2011). Comparative Genomics of 28 *Salmonella enterica* Isolates: Evidence for CRISPR-Mediated Adaptive Sublineage Evolution. *J Bacteriol.*, 14(193): 3556-3568.

Jackson B., [Bohin J. P.](#) & Kennedy E. P. (1983). Biosynthesis of Membrane-Derived Oligosaccharides: Characterization of *mdoB* Mutants Defective in Phosphoglycerol Transferase I Activity. *J Bacteriol.*, 160(3), 976-981.

Quirós Cárdenas Saúl (2016). Infecciones por bacterias del género *Salmonella*: Relevancia en la práctica clínica. *Rev. Clin. De EMED UCR.*

Schulman H. & Kennedy E. P. (1979). Localization of Membrane-Derived Oligosaccharides in the Outer Envelope of *Escherichia coli* and Their Occurrence in Other Gram-Negative Bacteria. *J Bacteriol.*, 137(1), 686-688.

Análisis bioinformático de una cisteín proteasa presente en plantas de *Nicotiana benthamiana*, la proteína Metacaspasa 4.

Ana Marchesini^{1,2}

¹ Cátedra de Bioinformática, Área de Biotecnología y Biología Molecular, Departamento de Ciencias Biológicas, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, La Plata, Argentina.

² Instituto de Biotecnología y Biología Molecular. Universidad Nacional de La Plata. Facultad de Ciencias Exactas - CONICET; Argentina.

RESUMEN

Los virus se encuentran dentro de los principales patógenos de los cultivos, causando pérdidas devastadoras, por lo que entender sus mecanismos de infección e interacción con la planta es fundamental para lograr su erradicación. Una de las formas es identificar las proteínas del hospedador partícipes en los mecanismos de defensa ante estos patógenos, como lo es la metacaspasa-4 (MCA4), una peptidasa que se activa ante estrés iniciando una reacción en cadena con el propósito de atacar al organismo invasor y evitar su propagación. En este trabajo se aborda un análisis bioinformático de la MCA4 presente en plantas *N. benthamiana* (NbMCA4) a fin de predecir su posible estructura e identificar segmentos fundamentales en su función, para poder inferir sobre su rol durante la infección viral.

PALABRAS CLAVE: Metacaspasa; *Nicotiana benthamiana*.

INTRODUCCIÓN

El crecimiento de la demanda de productos vegetales a nivel mundial, junto a la limitación de tierras cultivables, hacen cada vez más urgente la necesidad de mejorar las capacidades de producción en un marco de preservación del suelo y de sustentabilidad ambiental. Los virus están entre los principales patógenos de los cultivos, son responsables de pérdidas devastadoras en agricultura y dada la capacidad de adaptación a las estrategias de protección de los cultivos, su erradicación del ambiente parece imposible. Por esto, entender sus mecanismos de infección e interacción con la planta es fundamental para la producción de productos vegetales. Una de las características resaltantes de los virus es su capacidad de evadir los mecanismos de defensa de la planta, los cuales pueden ser receptores virales codificados por genes de resistencia (R) de la misma o mecanismos especializados, como PTGS y, para lograr un análisis profundo al respecto, primero se debe abordar el estudio de estos mecanismos.

Un grupo de proteínas, las metacaspasas (MCAs), funcionan como mecanismo primario de defensa de la planta ante el ataque de patógenos. Son cisteín proteasas representadas en todos los dominios de la vida (Uren et al., 2000), fueron descubiertas años más tarde que las caspasas a partir de una búsqueda de similitud estructural. De acuerdo a las preferencias de corte y estructura, las MCAs se han clasificado en 3 diferentes tipos, I, II y III, encontrándose sólo las MCAs tipo I y II en las plantas verdes (Viridiplantae). Ambos poseen el dominio catalítico conservado en las caspasas compuesto por las subunidades de 20 kDa (p20) y 10 kDa (p10); en la subunidad p20 se encuentran los residuos catalíticos Histidina-86 y Cisteína-139. Las MCAs tipo II poseen además una región de enlace de 160-180 aminoácidos entre p10 y p20. Estas proteínas se expresan como zimógenos y se activan por autoclivaje, esta actividad es regulada por modificaciones postraduccionales, como fosforilaciones, ubiquitinación, nitrosilación, cambios de pH y

concentración de iones, así como por la interacción con otras proteínas (Minina et al., 2017). Están involucradas en la muerte celular programada (PCD, *Programmed Cell Death*), programas de desarrollo y morfogénesis y en la contención de estreses bióticos y abióticos (He et al., 2008, Watanabe & Lam et al., 2011).

Nos propusimos identificar el rol que cumple la metacaspasa-4 presente en plantas de *Nicotiana benthamiana* (NbMCA4) durante la infección viral. La NbMCA4 es una metacaspasa tipo II, homóloga de la MCA4 de *A. thaliana*, la cual ha sido asociada a la activación de un programa de PCD bajo estrés oxidativo o ante el ataque de patógenos (Watanabe and Lam, 2011) a partir del clivaje en *trans* del PROPEP1 dando como producto PEP1, péptido que inicia una gran cadena de defensa en la planta (Hander et al., 2019). Con tal fin iniciamos un análisis de la biología de esta proteína, el cual se encuentra en actual estudio y muestra indicios de la importancia que podría tener en cuanto a la defensa de la planta. Nos planteamos entonces como objetivo de este trabajo realizar un estudio bioinformático de la NbMCA4 para poder recopilar mayor información en base a su secuencia, predecir su estructura y analizar su función en base a homologías, además de reconocer sitios potencialmente importantes para la misma, que permitirían el diseño de mutantes para fortalecer el análisis biológico.

MÉTODOS Y RESULTADOS

Análisis de secuencia

Para acceder a la secuencia de nucleótidos completa de NbMCA4 se utilizó la base de datos SolGenomics, ya que tanto en GenBank como en European Nucleotide Archive sólo se puede encontrar la anotación parcial (DQ084024), mientras que la secuencia primaria de la proteína no se encuentra anotada en forma completa en ninguna plataforma, la misma contiene una longitud de 418 aminoácidos, dando una proteína de 45,8 kDa de tamaño, y se proporciona a continuación

>NbMCA4

```
MAKKAVLIGINYPGTKAELKGCINDVKRMYSLIKRFGFSEEDITVLIDTDDSYTQPTGRNIRKVLSDLVGSAAEGDSL FVHYS  
GHGTRLPAETGEEDDTGYDECIVPCDMNLITDDDFRELVDKVPEGCRITIVSDSCHSGGLIDKAKEQIGESHKQGDDENEGH  
GSGFGFKFLRRSVEDAFESRGIHIPRRHDRREEEESFAESSVIETEDGDQVHVKNKSLPLSTLIEILKQKTGKDDIDVGKLRP  
TLFDVFGEDASPKVKKFMKVIFNKLQHGKGESEGGFLGMVGNLAQEFKQKLDENDESYAKPAMETHVEGKQEVYAGSG  
SRGLPDSGILVSGCQTDQTSADATPAGGDSYGALSNAIQEILAESDGPITNEEVTKARKKLQKQGYTQRPGLYCSDDHHVDA  
PFVC
```

A fin de analizar la estructura primaria de la NbMCA4, se procedió a una búsqueda de dominios en Pfam y en CDD (base de datos de dominios conservados, perteneciente a NCBI), encontrando en ambas plataformas un dominio conservado en su secuencia perteneciente a la superfamilia CASc en la región N-terminal de la proteína. Este dominio es típico de las caspasas, descrito en Pfam con el código de acceso 00656, llamado peptidase_C14. Se encuentran dentro de este los dos residuos conservados pertenecientes al sitio activo de las MCAs, H86 y C139 (subrayados en la secuencia proporcionada arriba).



Figura 1. Análisis en base a la secuencia primaria NbMCA4. Se observa la detección del dominio Peptidase_C14 perteneciente a la superfamilia de las CASc predicho por CDD.

Con el objetivo de inferir estructuras secundarias a partir de su secuencia, se utilizó el programa PSIPred, el cual asigna posibles plegamientos a las regiones e informa la confianza en la predicción. Se puede ver en la

Figura 2 que el arreglo aparentemente dominante serían las alfa-hélices, predichas con máxima confianza, unidas por una gran cantidad de loops. La misma búsqueda se realizó a partir del servidor Quick2D, el cual muestra predicciones de varios programas, incluido el PSIPred, sin indicar la confianza. Los resultados se muestran en el material suplementario (**Figura S1**) donde se puede ver que mayormente hay coincidencias entre ellos. En cuanto a la búsqueda de motivos o regiones coiled-coils a partir de la secuencia primaria, se corrieron los servidores Prosite y Coils, respectivamente, sin encontrarse nada en ninguno de los dos (datos no mostrados). Así mismo, se buscaron posibles segmentos transmembrana, para lo cual corrimos el servidor Phobius, donde se indicó la ausencia de un dominio de anclaje a lo largo de toda la cadena aminoacídica (dato no mostrado).



Figura 2. Predicción de estructuras secundarias a partir de la secuencia primaria utilizando el programa PSIPred. Las referencias por colores se observan en el extremo inferior.

Tanto la plataforma IUPred como DynaMine son muy usadas para la predicción de regiones desordenadas dentro de la secuencia proteica, se corrieron ambas, los resultados se muestran en la **Figuras 3a y 3b**. La mayor parte de la secuencia queda cerca del cutoff en ambos programas, lo cual podría ser coincidente con la gran cantidad de loops que se observaban en las predicciones de estructura secundaria, y aportan movimiento a la proteína. Además de segmentos cortos en los extremos, IUPred predice una región desordenada con un score levemente mayor al cutoff entre los residuos 150 y 170 aproximadamente, vista también en DynaMine junto con otros picos. Se ve en ambas plataformas segmentos con estructura muy ordenada, uno en la región N-terminal, abarcando aproximadamente del residuo 20 al 40, coincidente con la región donde se predice la presencia de una alfa-hélice de esa longitud, otro entre los aminoácidos 230-280, segmento en el cual se predice la presencia de 4 alfa-hélices unidas por loops muy cortos, y otro en el segmento 110-130.

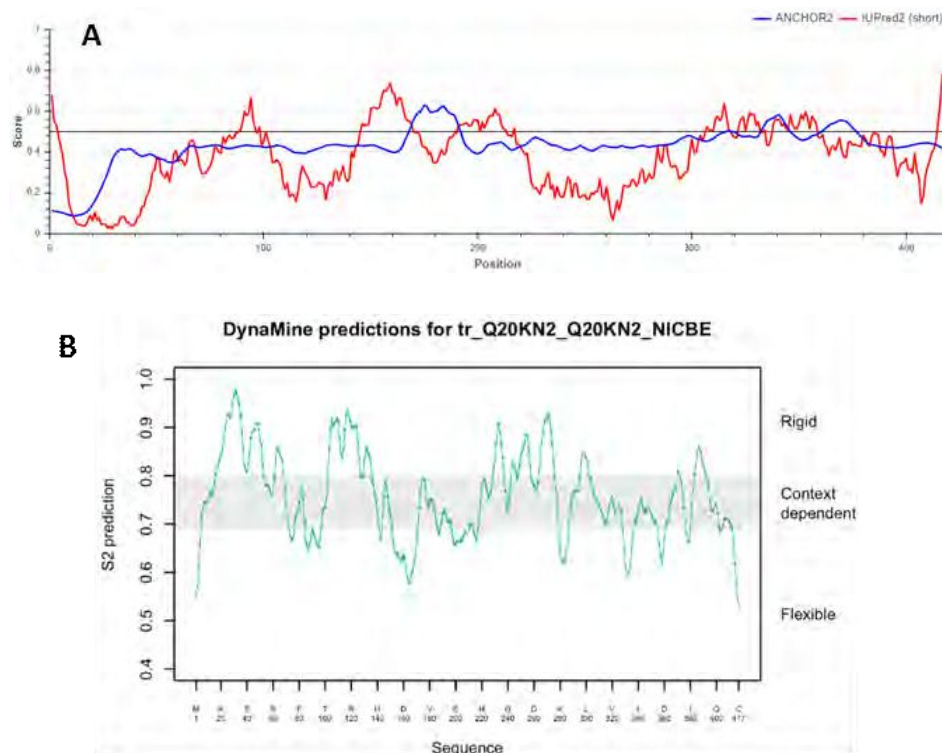


Figura 3. Predicción de regiones ordenadas/desordenadas a partir de la secuencia de NbMCA4 por los programas **a.** IUPred y **b.** DynaMine

Análisis de homología

Se buscaron secuencias homólogas usando el Blastp de NCBI para analizar la conservación de NbMCA4. Como base de datos se usó "non-redundant protein sequences", se indicó un máximo de 250 secuencias alineadas, un E-value de 0.05, ktuple de 6 aminoácidos y BLOSUM 62 como matriz. De un total de 250 proteínas encontradas todas presentaron más del 60% de similitud, con un *coverage* cercano o igual al máximo y valor de e-value 0.0. Se seleccionaron 81 homólogas pertenecientes al clado Mesangiosperma con cobertura de blast mayor o igual a 98% y se procedió a alinearlas a partir de su secuencia primaria.

En el material suplementario se puede ver el alineamiento resultante en formato MSF (**Figura S2**). Al hacer un análisis sobre éste encontramos que NbMCA4 comparte con todas las secuencias un porcentaje de identidad mayor o igual a 70% con gran cantidad de segmentos conservados y el número de gaps, dejando fuera del análisis a las escasas proteínas con distinta longitud, es muy bajo. La totalidad de las secuencias mantienen los residuos correspondientes al sitio activo H86 y C139 presentes en el dominio peptidase_C14. Estos datos muestran con seguridad que la MCA4 es una proteína muy conservada evolutivamente.

Modelado de la proteína: estructura terciaria

Para inferir la estructura terciaria, se realizó un modelado por homología. Como primer paso, la asignación de plegamiento, se utilizaron los programas HHpred, seleccionando como base de datos PDB_mmCIF70, y la plataforma de NCBI, BLASTp, usando como base de datos la PDB. En ambos se introdujo la secuencia primaria de NbMCA4 y los resultados mostraron como mejor template a la estructura cristalizada de la metacaspasa-4 de *A. thaliana* (AtMCA4), con un porcentaje de identidad del 70%, alto score y e-value cercano al 0 para HHpred y 0 para BLASTp, cubriendo el 100% de la secuencia. No hizo falta realizar un análisis con métodos más sensibles como psiBLAST ya que este porcentaje de similitud es más que suficiente para modelar la estructura terciaria de la NbMCA4 a partir del template, como así tampoco un análisis filogenético para evaluar cercanía evolutiva. En el caso de BLASTp, se encontró tanto la AtMCA4 wild type (wt) (código PDB: 6W8S) como su versión mutante C139A, donde se reemplazó la cisteína presente en el sitio activo por una adenina (código PDB: 6W8R), mientras que en HHpred sólo figura el mutante C139A. Continuamos nuestro análisis con la versión wt, 6W8S, ya que en 6W8R la mutación de un residuo tan importante podría afectar la estructura de la proteína a partir de la cual realizaremos el modelado.

Para corroborar la calidad del alineamiento a modo de asegurarnos un buen modelado se procedió a realizar el mismo en EMBOSS Needle (algoritmo Needleman-Wunsch), usando la matriz BLOSUM62 y analizarlo en Gendoc, obteniéndose un porcentaje de similitud del 81%, y 67% de identidad entre las secuencias, con sólo 23 gaps en todo el alineamiento. En la **Figura S3** del material suplementario se puede ver el mismo.

Una vez que validamos el alineamiento, se procedió a realizar una búsqueda en la base de datos PDB de todas de todas las estructuras cristalizadas para la proteína AtMCA4, encontrándose 6W8S con una resolución de 3.484 Å, y 6W8T con una resolución de 3.2 Å, esta última corresponde a la estructura vista cuando se tratan los microcristales con Ca²⁺; para comprender mejor esto nos remitimos a la publicación donde se reportaron las estructuras (Zhu et al., 2020) y vimos que la AtMCA4 se activa por clivaje en el residuo K225 en presencia de una concentración de Ca²⁺ elevada, y que la diferencia en las estructuras es que 6W8S corresponde a la proteína entera, presentando una distancia entre K225 y los residuos catalíticos C139 y H86 apta para el ataque nucleofílico, mientras que en la estructura 6W8T se puede ver que hubo un clivaje en el residuo K225 (las estructuras se muestran en la **Figura S4** del material suplementario). Decidimos proceder al análisis con la estructura del zimógeno, 6W8S.

Se pudo ver que la proteína AtMCA4 cristaliza como un monómero, con iones sulfato co-cristalizando con ella. Se identifica en ella un sólo dominio peptidase_C14 tal como habíamos predicho para la NbMCA4. Presenta una región desordenada (153-172) coincidente con los resultados arrojados por IUpred para NbMCA4.

Una vez encontrado el template, se descargó su secuencia en PDB, la cual tenía coordenadas desde el residuo 2 hasta el 420, presentando 64 missing residues, de los cuales 57 abarcan la región 152-208. Con estos datos se procedió a modelar a partir del programa Modeller, el cual genera distintas estructuras tratando de violar la menor cantidad de restricciones posibles. Se obtuvieron 25 modelos, siendo NbMCA4.B99990014 el más certero (**Figura 4**). Tal como se había predicho en su estructura secundaria, la NbMCA4 presenta una gran cantidad de segmentos hélices alfa, unas pocas beta plegadas y gran cantidad de loops.



Figura 4. Estructura obtenida a partir del modelado por homología de NbMCA4 usando el programa Modeller, con la estructura 6W8S de AtMCA4 como template.

Para evaluar el modelado se usó el programa ProSA, obteniéndose en la evaluación global un Zscore de -7.54, el cual cae dentro de la población de estructuras de referencia (**Figura 5a**). En cuanto a la evaluación local en el mismo programa, pudimos corroborar que el modelo es en su mayoría bueno, excepto por una corta región que presenta elevada energía, la cual podría corresponder al segmento de missing residues donde el programa tuvo que modelar *ab initio* (**Figura 5b**). En segundo lugar se realizó un alineamiento estructural entre el modelo obtenido y 6W8S a partir del programa TM -Align, donde se ve que el primero de estos muestra dos loops que no corresponden a la estructura de su homólogo (**Figura 5c**). El mismo análisis se realizó por el programa SuperPose, el cual muestra un RMSD total de 0.72, el cual es un valor muy aceptable de alineamiento estructural (dato no mostrado).

Análisis taxonómico

Para poder conocer el camino evolutivo de la proteína, se procedió a realizar un análisis taxonómico de esta utilizando el alineamiento múltiple anteriormente mostrado. Se utilizó el método de maximum likelihood realizado por el programa HyPhy para obtener el mejor modelo que explique la evolución de las secuencias, generando como primer árbol filogenético un Neighbor Joining, y asignando 4 categorías de velocidad de evolución para gamma distribution sea cual fuera el modelo de evolución resultante. El resultado arrojó que con un score de 31243.52409783448 y un likelihood de -15441.762, el mejor modelo es JTT+F. Con estos datos nos fuimos al programa PhyML para lograr estimar la topología de las secuencias alineadas a partir del modelo JTT+F+gamma, indicando que realice un análisis de bootstrapping no paramétrico (100 replicantes) para validar la información del alineamiento, el árbol resultante se puede ver en la **Figura 6** y a partir de él podemos inferir la cercanía evolutiva de NbMCA4 a las distintas homólogas, con un gran soporte en la información secuencial para los nodos resultantes.

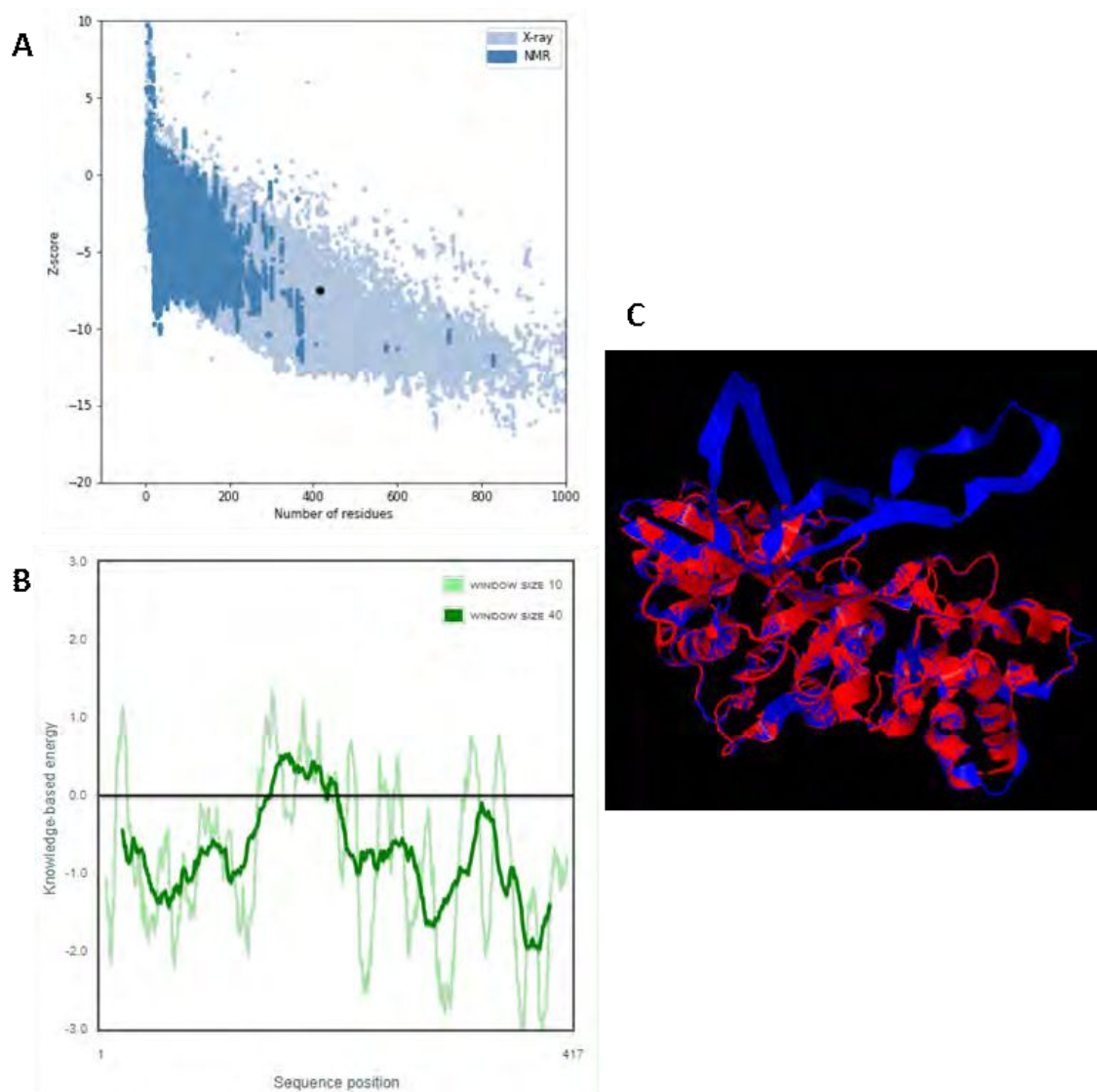


Figura 5. Evaluación del modelo estructural obtenido para NbMCA4. a y b. Evaluación energética global y local, respectivamente, usando el programa ProsaII; **c.** Alineamiento estructural entre el modelo obtenido y la estructura de referencia, 6W8S.

Predicción de la función

A la hora de predecir la función de una proteína de novo la única herramienta bioinformática que tenemos es la comparación con un homólogo cuya actividad esté reportada, y luego validar biológicamente esta información. En este caso, la proteína que vamos a utilizar es la AtMCA4.

La AtMCA4 está reportada como una cisteín proteasa que cliva específicamente luego de los residuos R o K, sin clivar sustratos específicos de las caspasas. Induce la muerte celular programada (PCD) al sensor estreses bióticos y abióticos, a partir del aumento de la concentración de Ca^{+2} intracelular. El sitio activo está compuesto por dos residuos, H86 y C139, los cuales están conservados en la NbMCA9, y cliva luego del residuo K225, el cual está presente en una región conservada en NbMCA4 como K224 (**Figura S3**), generando dos subunidades, p20 (1-225) y p10 (226-418). Sus anotaciones en Gene Ontology (GO) muestran que actúa en el citoplasma, citosol, mitocondria, membrana y plasmodesmos, funciona como una cistein-endopeptidasa con unión a proteínas presentando actividad hidrolasa y está involucrada como se había mencionado antes en la regulación de la PCD y de otros procesos biológicos involucrados en la defensa de la planta ante ataques bióticos y abióticos.

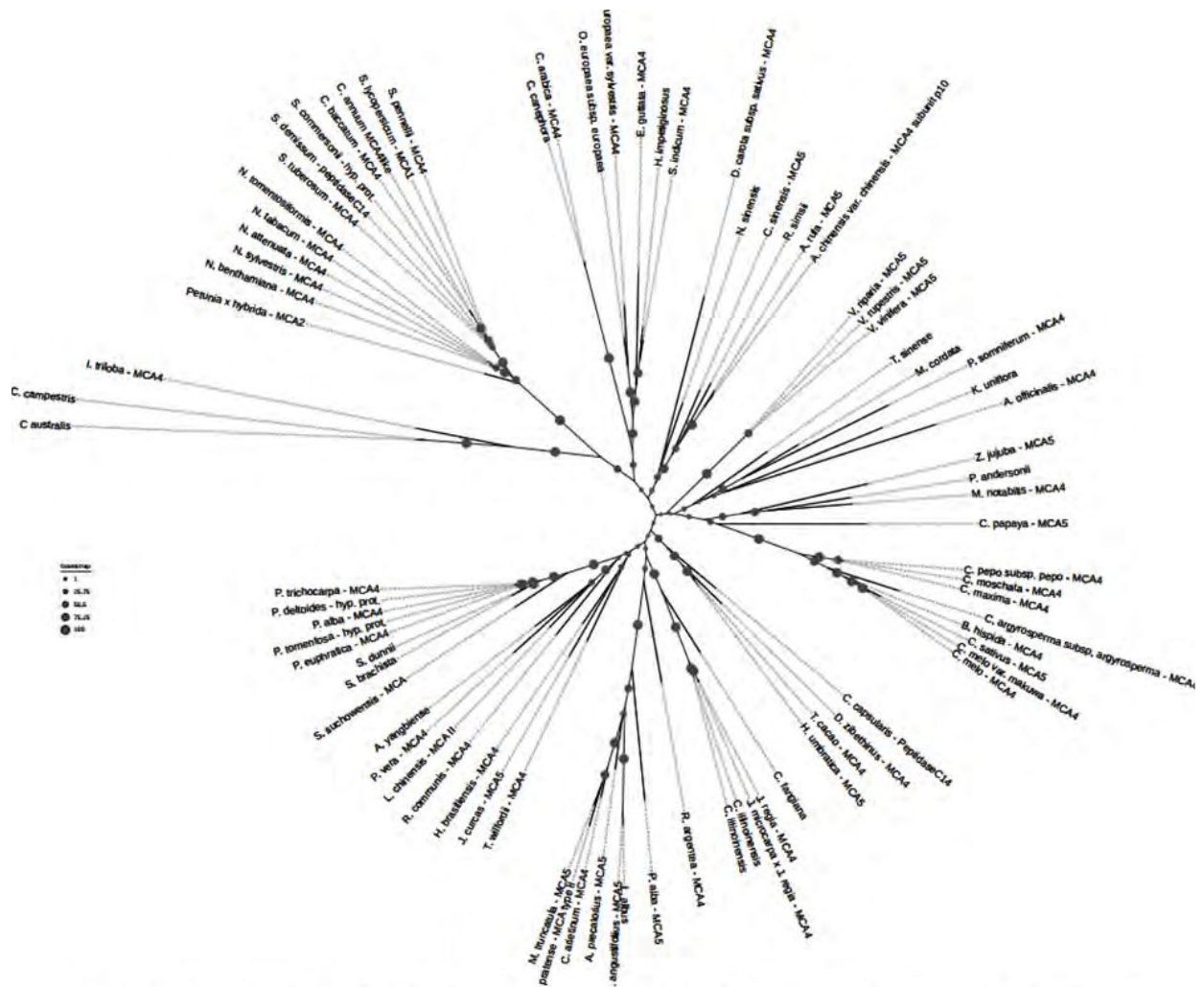


Figura 6. Árbol filogenético obtenido con el programa PhyML para 82 proteínas homólogas evolucionando a partir del modelo JTT+F+Gamma. Se muestra el soporte de cada nodo realizado a partir de un análisis de bootstrapping no paramétrico con 100 replicantes por el mismo programa.

A partir de la secuencia de NbMCA4 se corrió el programa Universal Evolutionary Trace (UET) y se vieron cuáles son los residuos más conservados evolutivamente, obteniéndose los resultados de la **Figura 7**, los cuales muestran una gran conservación a lo largo de toda la cadena, excepto por dos segmentos: 156-220 (región reportada con estructura desordenada en AtMCA4) y 267-338. Con esta misma secuencia pero agregando la estructura de referencia 6W8S cadena A (las 4 cadenas son idénticas) de AtMCA4 se inició el programa ConSurf y los resultados fueron similares: se conserva casi toda la secuencia exceptuando las mismas dos regiones (dato no mostrado).

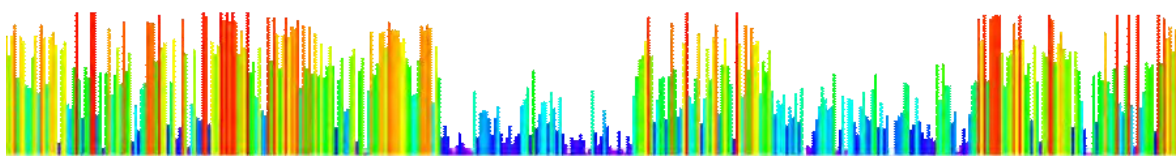


Figura 7. Análisis de residuos conservados evolutivamente en la MCA4 a partir del programa Universal Evolutionary Trace (UET). Los picos más altos y de color rojo corresponden a una mayor conservación, mientras que los de menor altura y azules a una menor conservación.

CONCLUSIONES Y DISCUSIÓN

La proteína NbMCA4 posee un dominio conservado en su secuencia perteneciente a la superfamilia CASC, el cual contiene los residuos pertenecientes al sitio activo, H86 y C139, también presentes en ella.

La estructura secundaria predicha de la NbMCA4 presenta mayormente arreglos alfa-hélice con algunas beta-plegadas y gran cantidad de loops.

Según las predicciones de IUPred, la NbMCA4 presenta regiones desordenadas en sus colas y en el segmento 150-170, este último reportado en la estructura de su proteína homóloga, AtMCA4.

NbMCA4 es una proteína muy representada a lo largo del clado mesoangiosperma, conservando la mayor parte de los segmentos de su secuencia, con dos regiones variables, una de ellas correspondiente a una posible estructura desordenada.

La proteína homóloga con estructura conocida más cercana es la AtMCA4, la cual presenta un 81% de identidad con NbMCA4 cubriendo el 100% de la proteína y su estructura es un monómero que co-cristaliza con iones sulfato (PDB: 6W8S).

A partir de la estructura de AtMCA4 se modeló la de NbMCA4, obteniendo buenos valores energéticos globales y locales, excepto por una región de elevada energía, y un alineamiento estructural bueno, con un RMSD de 0.72, donde se observa el modelado erróneo de dos loops. Para un mejor modelado habría que optimizar la región correspondiente a esos loops.

La NbMCA4 funcionaría como una cisteín proteasa, y se activaría auto-clivándose a partir de los residuos H86 y C139 luego del aminoácido K224 luego de sentir una concentración de Ca²⁺ intracelular elevada por un estrés biótico o abiótico sufrido por la planta hospedadora.

BIBLIOGRAFÍA

De Castro E, Sigrist CJA, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, Baiocchi A, Hulo N. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res.* 2006 Jul 1;34(Web Server issue):W362-5. PubMed:16845026 [Full text] [PDF version].

Di Tommaso P, Moretti S, Xenarios I, Orobítz M, Montanyola A, Chang JM, Taly JF, Notredame C. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res.* 2011 Jul;39(Web Server issue):W13-7. doi: 10.1093/nar/gkr245. Epub 2011 May 9. PMID: 21558174; PMCID: PMC3125728.

Drozdzetskiy A, Cole C, Procter J & Barton GJ. *Nucl. Acids Res.* (first published online April 16, 2015) doi: 10.1093/nar/gkv332

Elisa Cilia, Rita Pancsa, Peter Tompa, Tom Lenaerts, and Wim Vranken. The DynaMine webserver: predicting protein dynamics from sequence. *Nucleic Acid Research* doi: 10.1093/nar/gku270 (2014)

Gábor Erdős, Zsuzsanna Dosztányi. Analyzing Protein Disorder with IUPred2A. *Current Protocols in Bioinformatics* 2020;70(1):e99

GO Enrichment Analysis: Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* Jan 2019;47(D1):D419-D426.

Hander T, Fernández-Fernández ÁD, Kumpf RP, Willems P, Schatowitz H, Rombaut D, Staes A, Nolf J, Pottier R, Yao P, Gonçalves A, Pavić B, Boller T, Gevaert K, Van Breusegem F, Bartels S, Stael S. Damage on plants activates Ca²⁺-dependent metacaspases for release of immunomodulatory peptides. *Science.* 2019 Mar 22;363(6433):eaar7486. doi: 10.1126/science.aar7486. PMID: 30898901.

He R, Drury GE, Rotari VI, Gordon A, Willer M, Farzaneh T, Woltering EJ, Gallois P. Metacaspase-8 modulates programmed cell death induced by ultraviolet light and H₂O₂ in Arabidopsis. *J Biol Chem*. 2008 Jan 11;283(2):774-83. doi: 10.1074/jbc.M704185200. Epub 2007 Nov 12. PMID: 17998208.

Jon Ison, Matúš Kalaš, Inge Jonassen, Dan Bolser, Mahmut Uludag, Hamish McWilliam, James Malone, Rodrigo Lopez, Steve Pettifer, Peter Rice, EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats, *Bioinformatics*, Volume 29, Issue 10, 15 May 2013, Pages 1325–1332, <https://doi.org/10.1093/bioinformatics/btt113>

Kouza M, Faraggi E, Kolinski A, Kloczkowski A. The GOR Method of Protein Secondary Structure Prediction and Its Application as a Protein Aggregation Prediction Tool. *Methods Mol Biol*. 2017;1484:7-24. doi: 10.1007/978-1-4939-6406-2_2. PMID: 27787816.

Lu S et al. (2020). "CDD/SPARCLE: the conserved domain database in 2020.", *Nucleic Acids Res*. 48(D1):D265-D268.

Lukas Käll, Anders Krogh and Erik L. L. Sonnhammer. A Combined Transmembrane Topology and Signal Peptide Prediction Method. *Journal of Molecular Biology*, 338(5):1027-1036, May 2004

Madeira F, Park YM, Lee J, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*. 2019 Jul;47(W1):W636-W641. DOI: 10.1093/nar/gkz268.

Minina EA, Coll NS, Tuominen H, Bozhkov PV. Metacaspases versus caspases in development and cell fate regulation. *Cell Death Differ*. 2017;24(8):1314-1325. doi:10.1038/cdd.2017.18

New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Guindon S., Dufayard J.F., Lefort V., Anisimova M., Hordijk W., Gascuel O. *Systematic Biology*, 59(3):307-21, 2010.

Pfam: The protein families database in 2021: J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G.A. Salazar, E.L.L. Sonnhammer, S.C.E. Tosatto, L. Paladin, S. Raj, L.J. Richardson, R.D. Finn, A. Bateman. *Nucleic Acids Research* (2020) doi: 10.1093/nar/gkaa913

Sergei L Kosakovsky Pond, Art F Y Poon, Ryan Velazquez, Steven Weaver, N Lance Hepler, Ben Murrell, Stephen D Shank, Brittany Rife Magalis, Dave Bouvier, Anton Nekrutenko, Sadie Wisotsky, Stephanie J Spielman, Simon D W Frost, Spencer V Muse, HyPhy 2.5—A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies, *Molecular Biology and Evolution*, Volume 37, Issue 1, January 2020, Pages 295–299, <https://doi.org/10.1093/molbev/msz197>

SuperPose v1.0 (2004) Rajarshi Maiti, Gary Van Domselaar, Haiyan Zhang, and David Wishart

Uren AG, O'Rourke K, Aravind LA, Pisabarro MT, Seshagiri S, Koonin EV, Dixit VM. Identification of paracaspases and metacaspases: two ancient families of caspase-like proteins, one of which plays a key role in MALT lymphoma. *Mol Cell*. 2000 Oct;6(4):961-7. doi: 10.1016/s1097-2765(00)00094-0. PMID: 11090634.

Watanabe N, Lam E. Arabidopsis metacaspase 2d is a positive mediator of cell death induced during biotic and abiotic stresses. *Plant J*. 2011 Jun;66(6):969-82. doi: 10.1111/j.1365-313X.2011.04554.x. Epub 2011 Apr 28. PMID: 21395887.

Wiederstein & Sippl (2007). ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research* 35, W407-W410.

Y. Zhang, J. Skolnick, TM-align: A protein structure alignment algorithm based on TM-score, *Nucleic Acids Research*, 33: 2302-2309 (2005)

Zhu, P., Yu, XH., Wang, C. et al. Structural basis for Ca²⁺-dependent activation of a plant metacaspase. *Nat Commun* 11, 2249 (2020). <https://doi.org/10.1038/s41467-020-15830-8>

Estudio bioinformático de la proteína Aurora kinasa A de *Bos Taurus*

Josefina Ormaechea

Cátedra de Bioinformática, Área de Biotecnología y Biología Molecular, Departamento de Ciencias Biológicas, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, La Plata, Argentina.

RESUMEN

En la regulación del ciclo celular, las proteínas Aurora juegan un papel fundamental como reguladores de la segregación de cromosomas y la división celular. Debido a esto, su mal funcionamiento ha sido asociado al desarrollo de células malignas y hay interés en su utilización como blancos terapéuticos. En el presente trabajo se aplicaron diferentes herramientas bioinformáticas para obtener información secuencial y estructural de la proteína Aurora kinasa A, de la especie *Bos taurus* (Uniprot ID: Q2TA06). Esta presenta un dominio kinasa y una región N-ter desordenada. Además presenta un 88.8% de identidad con la proteína ortóloga humana, la cual se utilizó para construir el modelo de la proteína por homología. La estructura de la proteína permite observar un loop de activación y el sitio activo.

PALABRAS CLAVE: *Bos taurus*, Aurora kinasa A, bioinformática

INTRODUCCIÓN

La vaca o toro, cuyo nombre científico es *Bos taurus*, es una especie de mamífero artiodáctilo de la familia Bovidae, cuya alimentación es estrictamente herbívora. Estos mamíferos rumiantes fueron domesticados hace unos diez mil años en el Medio Oriente, y su ganadería se desarrolló progresivamente a lo largo y ancho de todo el planeta. Su importancia económica y ganadera deviene del consumo de su carne, leche y cuero. Argentina se posiciona mundialmente como el sexto productor mundial de carne vacuna, con una producción de 3,025 miles de toneladas en 2019, el 5% de la producción mundial¹.

Las proteínas de la familia Aurora kinasas cuentan con tres miembros en mamíferos: Aurora A, Aurora B y Aurora C. La proteína Aurora kinasa A (ARK-1) es una serina/treonina kinasa que participa en la regulación y progresión del ciclo celular. Se encuentra asociada al centrosoma y los microtúbulos del huso durante la mitosis y juega un rol importante en varios eventos mitóticos, incluyendo el establecimiento del huso mitótico, la duplicación del centrosoma, la separación del centrosoma, el alineamiento de los cromosomas, el check-point del ensamblado del huso mitótico y la citoquinesis.

La búsqueda de la proteína ARK-1 de *Bos taurus* en GenBank permite obtener la siguiente secuencia de 402 aminoácidos (N° acceso NP_001033117, Uniprot ID: Q2TA06), derivada del gen localizado en el cromosoma 13 codificado en 10 exones:

>Q2TA06

```
MDRCKENCISGPKTAVPLSDGPKRVPVAQQFPSQNPVSVNSGQAQRVLCPTNSSQRVPSQAQKLVSIQKPVQTLKQKPPQ
AASAPRPVTRPPSNTQKSKQPQPAPGNNPEKEVASKQKNEESKKRQWALEDFEIGRPLGKGFNGVYLAREKQSKFILALK
VLFKAQLEKAGVEHQLRREVEIQSHLRHPNILRLYGYPFHDAITRVYLILEYAPLGAVYRELQKLSKFDEQRTATYITELANALSYC
HSKRVIHRDIKPENLLLGSAGELKIADFGWSVHAPSSRRTTLCGTLDYLPPEMIEGRMHDEKVDLWSLGVLCYEFVGVGKPPFE
ADTYQETYRRISRVEFTFPDCVPEGARDLISRLKHNPSQRPTLKEVLEHPWIIANSKPSQCKKESTSKQS
```

Las siguientes características de la proteína ARK-1 en *Bos taurus* y su actividad biológica son inferidas de su homólogo humano, por similitud de secuencia. La proteína es necesaria para el correcto ensamblado del huso durante la mitosis y para la localización de NUMA1 y DCTN1 al

córtex celular durante la metafase. ARK-1 es también necesaria para la activación inicial de CDK1 en los centrosomas. Fosforila numerosas proteínas *targets*, como ARHGEF2, BORA, BRCA1, CDC25B, DLGP5, HDAC6, KIF2A, LATS2, NDEL1, PARD3, PLK1, RASSF1, p53/TP53, TPX2. Regula la actividad despolimerizadora de tubulina de KIF2A. Participa en la formación normal de axones, así como en la formación y estabilización de microtúbulos. ARK-1 es un regulador clave en la vía p53/TP53, particularmente en vías críticas en la transformación a células oncogénicas. Fosforila sus propios inhibidores. Regula los niveles de la proteína anti-apoptótica BIRC5 por supresión de la expresión del adaptador FBXL7 de la proteína ubiquitin-ligasa E3 mediante la fosforilación del factor de transcripción FOXP1.

La sobreexpresión de las proteínas Aurora kinasas ha sido descrita en varias células malignas, señalando su participación y rol como oncogenes en la génesis de tumores. Inhibidores de estas proteínas han sido propuestas y estudiadas como posibles terapias anticancerígenas (Yan et al., 2016). Además, su importancia en el desarrollo del ciclo celular ha sido estudiado en la maduración meiótica de oocitos en varias especies, tales como *Xenopus laevis*, *Mus musculus* y *Bos taurus*, brindando información acerca de los procesos bioquímicos involucrados en la formación de oocitos maduros.

El objetivo del trabajo es aplicar diferentes herramientas bioinformáticas para obtener información secuencial y estructural de la proteína Aurora kinasa A, de la especie *Bos taurus*.

MÉTODOS Y RESULTADOS

PRIMERA PARTE: Análisis secuencial

La búsqueda de la proteína ARK-1 de *Bos taurus* en UniProt permite obtener la información anotada para la misma en diferentes bases de datos. Se encuentran anotados los residuos 143, 162 y 274 como sitios de unión a ATP y el residuo 256 como aceptor de protón, perteneciente al sitio activo. También se encuentran anotadas las regiones 210-213 y 260-261 como regiones de unión a nucleótidos fosfato. Las anotaciones de Gene Ontology (GO) molecular function para la proteína ARK-1 son: unión a ATP, actividad serin treonin kinasa, unión a ubiquitina ligasa. Se encuentra regulada por modificaciones post traduccionales, fosforilación y desfosforilación, principalmente en el residuo Thr-288 a lo largo del ciclo celular.

Se buscó la proteína en la base de datos Pfam para evaluar la presencia de dominios. Los parámetros establecidos evalúan secuencias con un valor de corte de E-value de 1.0. Se encontró un dominio significativo, el dominio kinasa, desde el aminoácido 133 al 383, indicado en la figura 1. Se trata de un dominio que posee 12797 arquitecturas anotadas.



Figura 1. Esquema que muestra la predicción de dominios en la proteína ARK-1 de *Bos taurus* en Pfam. Se encontró un solo dominio estadísticamente significativo, el dominio kinasa, en la región 133 a 383.

La búsqueda de la presencia de segmentos transmembrana en el servidor TMHMM (<http://www.cbs.dtu.dk/services/TMHMM/>) no arrojó ningún segmento, lo que es esperable para la

proteína ya que se encuentra localizada en el centrosoma, el huso mitótico, centriolos y proyecciones neuronales, según las anotaciones en UniProt y de acuerdo a la función biológica de la proteína.

Utilizando el programa Parcoil Scoring Form en la página <http://cb.csail.mit.edu/cb/paircoil/cgi-bin/paircoil.cgi> se buscaron predicciones a partir de la secuencia de regiones con estructura terciaria coiled-coil, obteniendo como resultado que la proteína ARK-1 no presenta regiones con dicha estructura.

Se utilizó el programa IUPred2A (Prediction of Intrinsically Unstructured Proteins) en la página <https://iupred2a.elte.hu/> para buscar la presencia de regiones desestructuradas en la proteína, con los parámetros predeterminados (Long Disorder). El programa arrojó una región con gran probabilidad de carecer estructura en el N-ter de la proteína, y la región estructurada conteniendo el dominio kinasa anotado en Pfam entre los aminoácidos 133-383. En la figura 2 se muestra el gráfico obtenido del servidor.

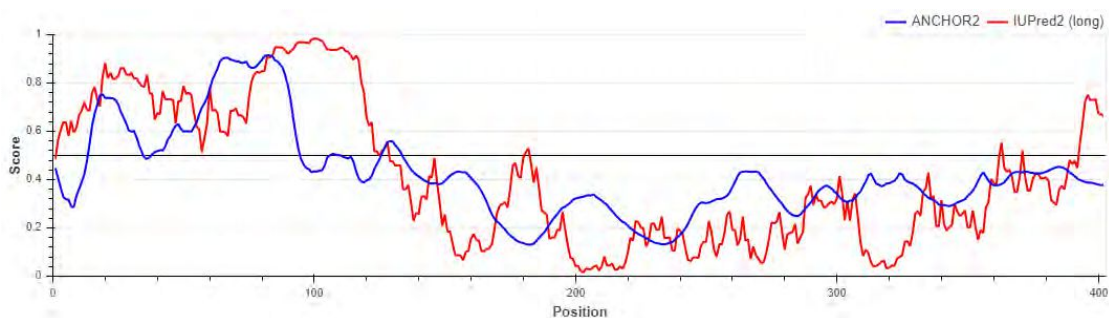


Figura 2. Predicción de segmentos desestructurados de la proteína ARK-1 de *Bos taurus* con el programa IUPred2A. Un score de uno se refiere a un residuo con alta probabilidad de ser desestructurado.

La búsqueda de predicción de la estructura secundaria en el servidor Quick 2D (<https://toolkit.tuebingen.mpg.de/tools/quick2d>) arroja un resultado concordante con la predicción de los segmentos desestructurados, con una región desordenada en el N-ter de la proteína. Luego, la predicción establece segmentos con α -hélice mayoritariamente, y algunas regiones cortas de cadenas β . Al utilizar el servidor PsiPred (<http://bioinf.cs.ucl.ac.uk/psipred/>) para la búsqueda de predicción de la estructura secundaria, el programa arroja un resultado similar en cuanto a las predicciones de α -hélice y cadenas β , pero con la gran diferencia de predecir en la región N-ter largas secuencias de coils. El programa incluye la confianza en la predicción de la estructura secundaria, y esta resulta baja para el N-ter, hasta aproximadamente el aminoácido 75, por lo que podría tratarse de una región desestructurada. El gráfico arrojado por el PsiPred se muestra en la figura 3.

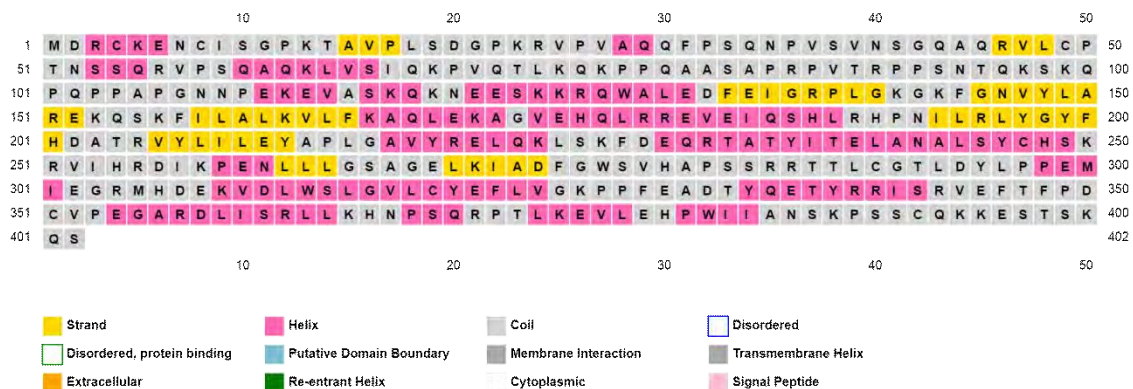


Figura 3. Figura que muestra la predicción de la estructura secundaria a partir del programa PsiPred. En rosa se encuentran los aminoácidos formando parte de una α -hélice y en amarillo aquellos involucrados en cadenas β . En gris, se predicen los residuos participantes en coils. La predicción hasta el aminoácido 75 posee baja confianza (no mostrado).

El análisis de la secuencia en MobiDB predice que el extremo N-ter de la proteína es desordenado, desde el aminoácido 1 al 126 (figura 4). El programa también otorga la predicción consenso del desorden, en la cual esta región no está incluida. Sin embargo, el MobiDB-lite es el único predictor que aparece, por lo que resulta raro que el consenso no incluya la región desordenada predicha por este.

Se realizó una búsqueda de proteínas homólogas utilizando BLAST en NCBI con los parámetros word size: 6, expect threshold: 0.05, la matriz BLOSUM62, gap penalty: 11 y gap extension: 1 y conditional compositional score matrix adjustment. Estos son los parámetros establecidos por el programa, y se utilizó para recuperar 5000 secuencias de una base de datos de proteínas no redundante. El programa recuperó 5000 secuencias de la base de datos. Se aplicó un filtro de coverage de 70% al 100% para obtener homólogos de la proteína completa. Con este filtro, el programa devuelve 2445 secuencias de las 5000 previas.

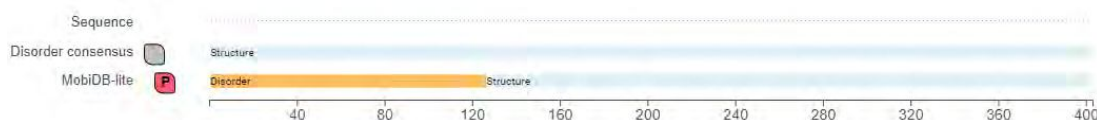


Figura 4. Imagen de la predicción de segmentos desordenados en el programa MobiDB.

La secuencia recuperada con menor porcentaje de identidad es de 42.35% de identidad. La secuencia con mayor valor de E-value es de $5e-101$. Por lo tanto considero que las secuencias recuperadas corresponden a proteínas homólogas cercanas. Todas pertenecen al dominio eucariota, y se encontraron homólogos en estramenopilos, algas verdes, amebas, algas rojas, hongos y metazoos.

Se realizó luego un PSI-BLAST utilizando los parámetros predeterminados, iguales a los parámetros del BLAST salvo por un word size de 3, para recuperar 1000 secuencias. Se aplicó un filtro de coverage a partir del 70%, y con las 919 proteínas se realizó un Profile y una búsqueda de 1500 secuencias con un valor umbral de E-value de 0.005. El programa devolvió 1342 secuencias. Al realizar otra iteración del nuevo profile construido frente a la base de datos de proteínas no redundantes y solicitar la recuperación de 5000 secuencias, el programa devuelve 3789 secuencias con coverage mayor al 70%. Si comparamos este valor con las 2445 secuencias halladas con el BLAST, evaluamos que se recuperaron más proteínas homólogas. Sin embargo, el porcentaje de identidad menor encontrado es de 48.96%, por lo que no se recuperaron homólogos lejanos. Por otro lado, la realización del primer profile con 919 proteínas puede haber introducido un sesgo hacia las proteínas más similares.

Para evaluar homólogos más lejanos, se realizó un PSI-BLAST contra una base de datos de proteínas clustereadas, lo que permite crear perfiles con mayor diversidad y por lo tanto más representativos de las diferencias entre las proteínas homólogas. Se utilizó el programa en MPI Bioinformatics Toolkit, la base de datos clustereada no redundante nr50_12_Apr predeterminada por el sitio, la matriz BLOSUM62, 3 iteraciones, un umbral de E-value para la inclusión en los perfiles de $1e-8$, un umbral de E-value para el reporte de resultados de $1e-3$ y 5000 como máximo de secuencias recuperadas. El programa recuperó 5000 secuencias de la base de datos. El E-value mayor es de $8.18e-109$ y corresponde a la secuencia titulada "hypothetical protein FGO68_gene235 [*Halteria grandinella*]", la especie corresponde a un protozoo ciliado.

Para evaluar la existencia de homólogos fuera del dominio Eukarya se realizó un BLAST en la página del NCBI con una base de datos de proteínas no redundante y se excluyeron aquellas secuencias en el dominio eucariota. Se utilizaron los parámetros predeterminados detallados previamente, y se realizó la búsqueda para 1000 secuencias. Al filtrarlas para un coverage mayor a 70%, el programa devolvió 69 secuencias. De estas secuencias, solo 15 pertenecen a organismos, mientras que el resto son construcciones sintéticas.

Todas las secuencias derivan de ensayos de ensamblado de metagenoma (MAG: Metagenome assembled genome) a partir de diversas muestras. Las secuencias son entonces proteínas hipotéticas anotadas a partir de bases de datos de proteínas procariotas, y figuran pertenecientes a los siguientes géneros u especies: *Clostridia*, *Waddilaceae*, *Planctomycetes*, *Candidatus Dadabacteria*, *Candidatus Riflebacteria*, *Thermoleophilia*, *Candidatus Latscibacteria*, *Candidatus Brocadiae*, *Gemmatimonadetes*, *Acidobacteria*, *Candidatus Eisenbacteria*.

Con el objetivo de evaluar el alineamiento de proteínas homólogas a ARK-1 a lo largo del árbol de la vida y los residuos conservados entre ellas se eligieron secuencias distribuidas en diferentes ramas del árbol y se preparó un archivo de texto del tipo FASTA para realizar posteriormente un análisis de alineamiento múltiple global. Se escogieron al azar las siguientes secuencias con más del 70% de coverage: 5 secuencias provenientes de bacterias, 2 secuencias de algas verdes y 2 de algas rojas, 6 secuencias del reino Plantae, 6 secuencias del reino Fungi y 10 secuencias pertenecientes a animales. Se realizó un alineamiento múltiple de secuencia en el programa T-Coffee. El alineamiento muestra que entre todas las secuencias, el dominio kinasa se encuentra altamente conservado mientras que las regiones N-ter y C-ter de las proteínas poseen mayores divergencias. Las proteínas alineadas probablemente no posean la misma función biológica, especialmente aquellas obtenidas de bacterias, para cuyas secuencias no está anotada la función biológica. Para las secuencias obtenidas de animales, la similitud es mayor y mejor el alineamiento. El dominio kinasa, presente en organismos tan divergentes como bacterias y mamíferos, implica su temprana aparición en el desarrollo de la vida. Por lo tanto se podrían estar recuperando proteínas que no pertenecen a la subfamilia de Aurora. En la figura 5 se muestra un fragmento del inicio de la región del dominio kinasa alineada por el T-coffee.

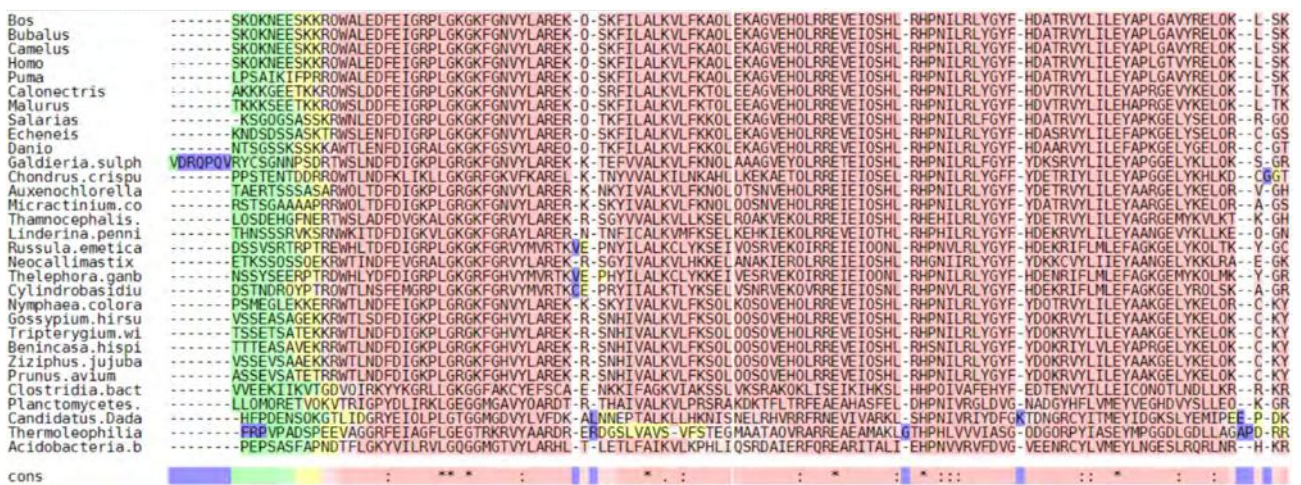


Figura 5. Alineamiento múltiple de secuencia en T-coffee, incluyendo secuencias de bacterias, animales, plantas, hongos y algas.

SEGUNDA PARTE: Análisis estructural y filogenético

Se escogió la proteína 1MQ4, código PDB correspondiente a la estructura cristalizada de la proteína Aurora A de *Homo sapiens* como template para el modelado por homología de la proteína Aurora A de *Bos taurus*. La proteína humana de 403 aminoácidos presenta un 88.8% de identidad con la proteína ortóloga bovina (358/403 aminoácidos). La estructura PDB elegida como molde no presenta mutaciones, fue cristalizada por el método de difracción de rayos X y posee una resolución de 1.90 Å (Nowakowski et al., 2002). La estructura cristalizada de la proteína abarca desde el aminoácido 126 al 388, que incluye el dominio kinasa. De los 263 aminoácidos cristalizados, el porcentaje de identidad es de 96,6% (254/263). Se incluye el alineamiento entre ambas secuencias en la figura 6.

```

Bos taurus 126 RQWALEDFEIGRPLGKGFQNVYLAREKQSKFILALKVLFKAQLEKAGVEHQLRREVEIQSHLRHPNILRLYG 198
Homo sapiens 126 RQWALEDFEIGRPLGKGFQNVYLAREKQSKFILALKVLFKAQLEKAGVEHQLRREVEIQSHLRHPNILRLYG 198

Bos taurus YFHDA TRVYLILEYAPLGA VYRELQKLSKFDEQRTATYITELANALSYCHSKRVIHRDIK PENLLLGSAGELK 271
Homo sapiens YFHDA TRVYLILEYAPLGA VYRELQKLSKFDEQRTATYITELANALSYCHSKRVIHRDIK PENLLLGSAGELK 271

Bos taurus IADFGWSVHAPSSRRRTTLCGTLDYLPPEMIEGRMHDEKVDLWSLGVLCYEFVLVGGKPPFEADTYQETYRRISR 344
Homo sapiens IADFGWSVHAPSSRRRTTLCGTLDYLPPEMIEGRMHDEKVDLWSLGVLCYEFVLVGGKPPFEADTYQETYKRISR 344

Bos taurus EFTFPDCVPEGARDLISRLLKHNPSQRPTLKEVLEHPWITAN - S 387
Homo sapiens EFTFPDFVTEGARDLISRLLKHNPSQRPMLEVLEHPWITANSS 388

```

Figura 6. Alineamiento entre la secuencia de aminoácidos cristalizada del gen Aurora A de *Bos taurus* y *Homo sapiens*. Se indica con una flecha las posiciones no conservadas.

El modelado por homología se realizó por el programa Modeller a partir del alineamiento entre ambas secuencias proteicas. La calidad del modelo se analizó por el DOPE score, obteniendo el mejor modelo entre los 10 generados el valor de -33556.46094. El análisis del modelo en el servidor PROSA generó los gráficos mostrados en la imagen 7 A) y B), indicativos de la calidad del modelo. El Z-Score del modelo es de -6.8, indicando que se encuentra dentro del rango para modelos de proteínas del mismo tamaño en la base de datos PDB. En el gráfico de energía por residuo se observa un fragmento que posee más energía. Al graficar el valor DOPE por posición del modelo se observa un pico en la misma región, los residuos con valores más positivos que -0,020 corresponden a las posiciones 282-289 (figura 7 C).

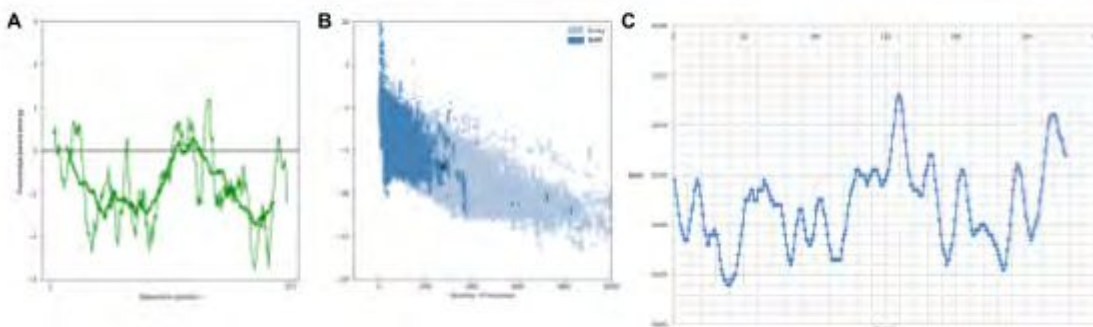


Figura 7. Figuras que indican la calidad del modelo estructural. A) Gráfico generado en PROSA de la energía de cada residuo por posición. Se observa una región levemente positiva. B) Gráfico que grafica el Z-score del modelo en un gráfico conteniendo los Z-scores de los modelos contenidos en PDB. El modelo generado se encuentra dentro de los valores esperados para modelos de la misma longitud. C) Gráfico del DOPE score por posición. Se observa un pico en la misma región del gráfico A, entre las posiciones 282 y 289. Además se observa un pico en el extremo C-ter de la proteína.

Al visualizar el modelo (figura 8) se observa que la región de mayor energía corresponde al loop de activación (Nowakowski et al., 2002).

Se utilizó el programa MEGA X para realizar un árbol filogenético entre varias secuencias de proteínas de diversos organismos recuperadas a partir de un BLAST contra la base de datos del NCBI. Además se eligieron aquellas secuencias anotadas como Aurora A y que representarán una amplia variedad de organismos. El alineamiento múltiple por el algoritmo MUSCLE revela que el dominio kinasa se encuentra bien conservado y el N-ter varía entre las especies, observándose grandes regiones con gaps. Por lo tanto, el N-ter se excluyó para el análisis filogenético, utilizando la región más conservada de la proteína para el análisis posterior, siguiendo el trabajo realizado por Brown et al (2004). Se evaluaron diferentes modelos de evolución de proteínas para el alineamiento realizado por MUSCLE por el método de Maximum Likelihood (ML). El modelo LG+G resultó el mejor modelo, dado por el BIC Score más bajo. Luego se estimó la filogenia utilizando el método ML con el modelo óptimo LG+G con 4 categorías discretas de velocidad de cambio

evolutivo. Se muestra en la figura 9 el árbol inferido a partir de 150 réplicas por el método de Bootstrap no paramétrico.

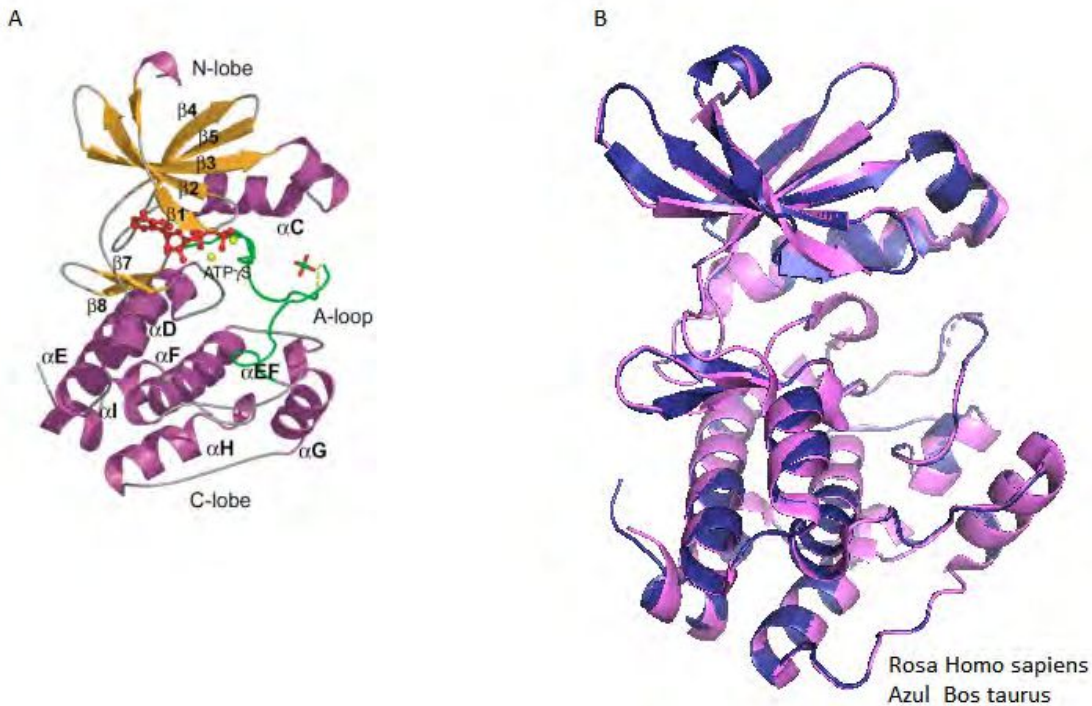


Figura 8. A) Estructura de Aurora kinasa A de *Homo sapiens*. Se indican los α hélices en rosa, las hojas plegadas β en amarillo, el loop de activación en verde, los iones Mg^{2+} como esferas amarillas y el fosfato unido rojo y verde. Imagen obtenida y adaptada de Nowakowski et al. 2002. B) Imagen del modelo generado de la proteína Aurora kinasa A de *Bos taurus* y la estructura molde de la misma proteína de *Homo sapiens* alineados en PyMOL.

Se observa que en el árbol hay dos grupos separados, uno correspondiente a vertebrados, y otro en el que se agrupan secuencias pertenecientes a plantas e invertebrados. Los hongos quedan agrupados, evolutivamente más alejados del resto. Los genes en el cluster de vertebrados son ortólogos. La baja confianza en las ramas del otro cluster no permiten afirmar sus relaciones evolutivas con certeza. Se resalta una divergencia evolutiva entre las secuencias de ambos grupos. Los resultados son concordantes con los descritos por Brown et al (2004). En este trabajo se menciona que los invertebrados poseen dos proteínas homólogas Auroras, los hongos incluidos en el presente alineamiento tienen una única proteína Aurora, y los mamíferos tres. Además concluyen que las proteínas Aurora A de mamíferos y vertebrados de sangre fría son ortólogas, mientras que las proteínas Aurora B y Aurora C de mamíferos evolucionaron más recientemente a partir de un evento de duplicación en vertebrados de sangre fría.

El análisis del modelo pdb de Aurora A de *Bos taurus* por el programa Evolutionary Trace resulta en la coloración de los residuos según la predicción de su importancia biológica tanto en la secuencia como en el modelo 3D de la proteína. Se observa que la región correspondiente al loop de activación y el túnel se encuentran predichos como regiones de mayor importancia. Se incluye en la figura 10 una imagen de la superficie del modelo generado en PyMOL en donde se resaltó en color magenta el loop de activación, el cual dio la energía más alta por DOPE previamente. Por otro lado, se señalan en naranja los 26 aminoácidos indicados por Brown et al. (2004) que se encuentran en el sitio activo de la proteína Aurora A de *Homo sapiens*. De estos residuos, uno solo difiere entre los ortólogos humano y vacuno, indicado en verde, que corresponde a un cambio de T por A en la posición 217 de la secuencia de *Bos taurus*.

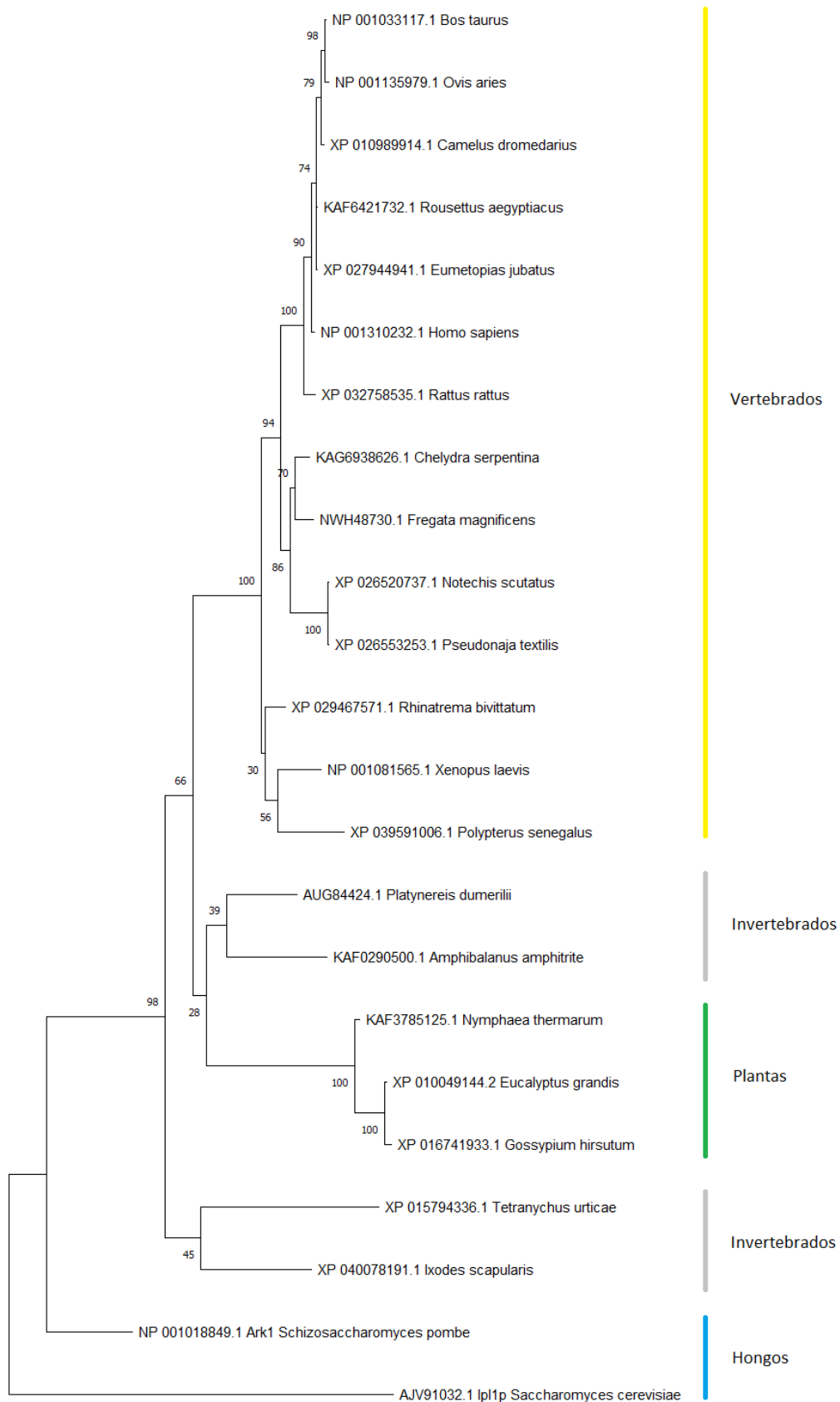


Figura 9. Árbol filogenético inferido a partir del alineamiento múltiple del dominio conservado kinasa de distintas secuencias homólogas de proteínas de Aurora A. Se utilizó el método de Maximum Likelihood con el modelo LG más la distribución gamma con 4 categorías discretas de velocidad de evolución. El árbol inicial fue obtenido automáticamente por los algoritmos Neighbor Joining y BioNJ. Se muestra el árbol a partir de 150 replicados por el método Bootstrap. El análisis involucró 23 secuencias, de las que se observa la especie de la cual derivan y el código de acceso del NCBI.

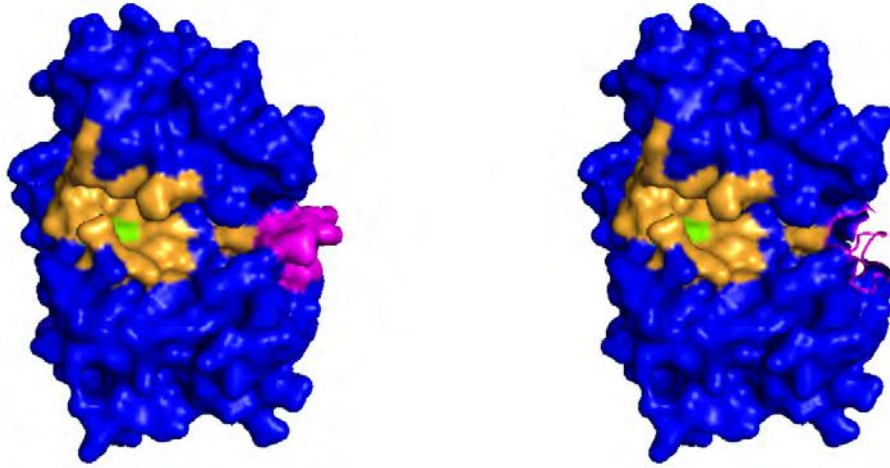


Figura 10. Imágenes de la superficie del modelo generado para el dominio funcional de Aurora A de *Bos taurus* en PyMOL. Se indica en magenta el loop de mayor energía, correspondiente al loop activador. En naranja se resaltan los residuos lindantes al sitio activo para el homólogo humano, según Brown et al (2004). En verde se indica el residuo no conservado entre los homólogos humano y vacuno, que corresponde a un cambio de T por A en la posición 217 de la secuencia para *Bos taurus*.

Al evaluar los mismos 26 residuos lindantes al sitio activo en el MSA utilizado para la construcción del árbol filogenético, se observa que 14 de las 26 posiciones señaladas no se encuentran conservadas para todas las secuencias (figura 11).

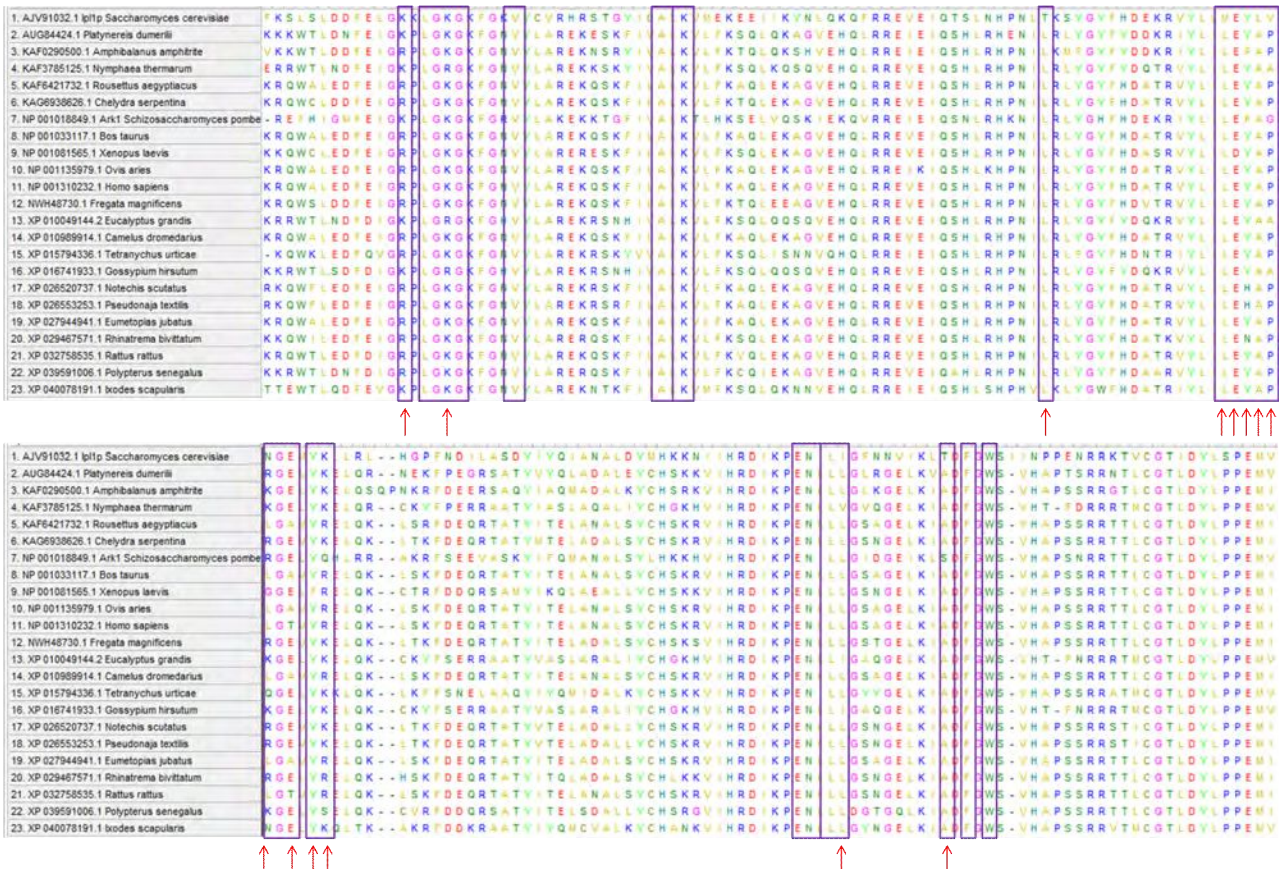


Figura 11. Se muestra una parte del alineamiento múltiple de las secuencias incluidas en el armado del árbol filogenético. Se resaltan con violeta los sitios lindantes al sitio activo según Brown et al 2004, y con una flecha roja aquellos que no están conservados para todas las secuencias. De los 26 residuos, 14 no son idénticos.

CONCLUSIONES Y DISCUSIÓN

La proteína Aurora A de *Bos taurus* presenta un alto porcentaje de identidad (88.8%) con su ortólogo humano. Esta proteína homóloga se encuentra cristalizada y ampliamente estudiada como posible target de drogas anti tumorales.

Aurora A bovina presenta un segmento N-ter desestructurado y un dominio kinasa seguido de una corta región de aminoácidos en el extremo C-ter.

El análisis filogenético reveló que las proteínas Auroras A en vertebrados son ortólogas. Estas presentan diferencias evolutivas con sus homólogas en invertebrados, las cuales se agrupan en el mismo cluster que las secuencias de hongos y plantas.

Con respecto al MSA, resulta llamativo que un alto porcentaje de residuos lindantes al sitio activo no se encuentren conservados entre los distintos organismos (14/26, 54%). Por otro lado, el dominio N-ter de las proteínas es desestructurado y especie específico. Sería interesante realizar un análisis de las regiones N-ter, y un análisis estructural entre el dominio kinasa de proteínas con diferente función biológica versus el dominio kinasa de las proteínas Aurora, de manera de dilucidar relaciones funcionales en estos dominios ampliamente distribuidos.

BIBLIOGRAFÍA

Franco Ramseyer - Emilce Terré (08 de noviembre de 2019) Carne vacuna en el mundo en niveles récord, Argentina aprovecha la mayor demanda global. Bolsa de Comercio de Rosario. <https://www.bcr.com.ar/es/mercados/investigacion-y-desarrollo/informativosemanal/noticias-informativo-semanal/carne-vacuna>

Yan M, Wang C, He B, Yang M, Tong M, Long Z, Liu B, Peng F, Xu L, Zhang Y, Liang D, Lei H, Subrata S, Kelley KW, Lam EW, Jin B, Liu Q. (2016) Aurora-A Kinase: A Potent Oncogene and Target for Cancer Therapy. *Med Res Rev.* 2016 Nov;36(6):1036-1079. doi: 10.1002/med.21399.

J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G.A. Salazar, E.L.L. Sonnhammer, S.C.E. Tosatto, L. Paladin, S. Raj, L.J. Richardson, R.D. Finn, A. Bateman. (2020) *Pfam: The protein families database in 2021*. *Nucleic Acids Research*. doi: 10.1093/nar/gkaa913

Bonnie Berger, David B. Wilson, Ethan Wolf, Theodore Tonchev, Mari Milla, and Peter S. Kim, (1995) "Predicting Coiled Coils by Use of Pairwise Residue Correlations", *Proceedings of the National Academy of Science USA*, vol 92, pp. 8259-8263.

Bálint Mészáros, Gábor Erdős, Zsuzsanna Dosztányi. (2018) [IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding](#) *Nucleic Acids Research*; 46(W1):W329-W337.

Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V. (2018) A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. [J Mol Biol. S0022-2836\(17\)30587-9](#).

Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Res.* 25:3389-3402.

Gabler F, Nam SZ, Till S, Mirdita M, Steinegger M, Söding J, Lupas AN, Alva V. (2020) Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. [Curr Protoc Bioinformatics. 72\(1\):e108](#). doi: 10.1002/cpbi.108.

Notredame, Higgins, Heringa. (2000) T-Coffee: A novel method for multiple sequence alignments. *JMB*, 302 (205-217)

Nowakowski, J., Cronin, C. N., McRee, D. E., Knuth, M. W., Nelson, C. G., Pavletich, N. P., Thompson, D. A. (2002). Structures of the Cancer-Related Aurora-A, FAK, and EphA2 Protein Kinases from Nanovolume Crystallography. *Structure*, 10(12), 1659–1667. doi:10.1016/s0969-2126(02)00907-3

Markus Wiederstein, Manfred J. Sippl. (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins, *Nucleic Acids Research*, Volume 35, Pages W407–W410

Kumar S., Stecher G., Li M., Knyaz C., and Tamura K. (2018) MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. Molecular Biology and Evolution 35:1547-1549.

Brown, J.R., Koretke, K.K., Birkeland, M.L. et al. (2004) Evolutionary relationships of Aurora kinases: Implications for model organism studies and the development of anti-cancer drugs. BMC Evol Biol 4, 39

Modelado molecular y predicción funcional de la proteína bll4781 de *Bradyrhizobium diazoefficiens* (USDA 110)

Damián Brignoli

Cátedra de Bioinformática, Área de Biotecnología y Biología Molecular, Departamento de Ciencias Biológicas, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, La Plata, Argentina.

RESUMEN

La proteína denominada bll4781 pertenece a la bacteria de suelo *Bradyrhizobium diazoefficiens* (USDA 110), rizobio de importancia agronómica por su capacidad de fijar N atmosférico en las raíces de leguminosas. La secuencia de esta proteína presenta una longitud de 100 aa, no presenta regiones transmembrana y presenta regiones desordenadas entre las posiciones 1-15 y 83-100, respectivamente. Se realizó una búsqueda de homología por diversos métodos y programas, y se encontraron homólogos remotos con estructura conocida, y sólo uno de ellos presentó la mayor significancia estadística, una proteína piruvato dehidrogenasa perteneciente a la bacteria fotosintética *Rhodospseudomonas palustris* CGA009. Las relaciones filogenéticas muestran mayor cercanía evolutiva con varias especies de *Bradyrhizobium*, formando un clado de 8 especies donde se encuentra el representante de la proteína de estructura conocida. Para predecir la posible función biológica de la proteína se realizó una búsqueda exhaustiva en diferentes bases de datos realizando el seguimiento del nº EC, encontrando que la proteína de estudio participa en las rutas metabólicas de la fermentación del acetato y metabolismo del piruvato, realizando la posible función molecular de catalizar la reacción entre el piruvato y una molécula de ubiquinona, dando como resultado acetato, ubiquinol y dióxido de carbono.

PALABRAS CLAVE: piruvato deshidrogenasa, función, bioinformática

INTRODUCCIÓN

La proteína denominada bll4781 pertenece a la bacteria de suelo *Bradyrhizobium diazoefficiens* (USDA 110), rizobio Gram negativo capaz de fijar el N atmosférico en órganos especializados de las raíces de las leguminosas denominados nódulos. La simbiosis rizobio-leguminosa es de gran importancia agronómica y permite que la planta crezca con éxito en ausencia de fertilizantes nitrogenados suministrados de forma externa. Esta bacteria, junto a otras de su misma especie, como por ejemplo *Bradyrhizobium elkanii* y *Bradyrhizobium japonicum* (referidas conjuntamente como *Bradyrhizobium spp.*), forman parte del grupo de microorganismos conocidos como PGPR (Plant growth-promoting rhizobacteria) debido a la propiedad mencionada anteriormente. Gracias a esta propiedad, se utilizan como componentes principales de los inoculantes para soja, los cuales hoy en día son ampliamente utilizados y la práctica de la inoculación está ampliamente extendida.

El genoma de esta bacteria es un cromosoma circular de 9.105.828 bp de longitud que comprende 8317 posibles genes codificantes de proteínas, de los cuales el 52% muestra una secuencia similar a genes de función conocida y un 30% a genes hipotéticos. El restante 18% no muestra una similitud aparente a los restantes genes reportados.

El objetivo general de este trabajo es predecir la posible función biológica de la proteína en cuestión a partir de conocer su posible estructura utilizando algunas de las herramientas brindadas durante el curso.

MÉTODOS Y RESULTADOS

En particular, la secuencia de aminoácidos de la proteína bll4781 tiene una longitud de 100 aa. A continuación se muestra la secuencia de aa de dicha proteína en formato FASTA:

```
>tr|Q89KW9|Q89KW9_BRADU Bll4781 protein OS=Bradyrhizobium diazoefficiens (strain JCM 10833 / BCRC 13528 / IAM 13628 / NBRC 14792 / USDA 110) OX=224911 GN=bll4781 PE=4 SV=1
```

```
MAGPKEQLPPDVVTREDAVEILRVFVLDGGLSMAFQRAFEEDMWGLLLVDLARHAARAYARESEYTEEDALSRIEMFQA  
EIERPTDTGTTTPRGKGH
```

En la búsqueda de homólogos cercanos y remotos se hizo uso de los programas BLAST y PSI-BLAST, de la base de datos del NCBI (National Center for Biotechnology Information), así como de los programas HMMER y HHpred, del servidor Toolkit.

Se realizó una primera búsqueda con BLAST con un parámetro de 1000 secuencias iniciales, descartando aquellas secuencias con porcentajes de identidad menores al 30%, con lo cual quedaron establecidas 521 secuencias como posibles homólogos cercanos. A continuación, se realizó una búsqueda de homólogos remotos por PSI-BLAST, también con un parámetro inicial de 1000 secuencias, arrojando en la primera iteración un resultado de 765 secuencias, luego se realizaron 3 iteraciones más donde el programa encuentra y añade secuencias con menos del 30% de identidad a partir de la tercera iteración. Para un análisis de mayor sensibilidad, se utilizó el programa JACKHMMER (secuencia-profile) del EMBL-EBI, encontrando 463 "matches" con significancia estadística en la primera iteración. La búsqueda se realizó por base de datos de UniProt, con restricción por taxonomía (Eubacteria) y sin modificación de parámetros numéricos preestablecidos. Por último, se utilizó el servidor Toolkit para correr el programa HHpred seleccionando la base de datos del PDB, con el cual llegamos al grado de mayor sensibilidad (profile-profile) en la búsqueda de homología remota. El programa encontró 27 "hits", de los cuales el primero de ellos tuvo la mayor significancia estadística (E-value 7,9 e-35), Score (192,7) y porcentaje de Identidad (%76).

La distribución taxonómica de los resultados obtenidos por BLAST y PSI-BLAST muestra que el mayor número de microorganismos representados corresponden a proteobacterias, más precisamente a alfa-proteobacterias, seguidas de gamma y beta-proteobacterias, encontrándose como más distantes a organismos como *Calditrichaeta bacterium* y un nemátode (*Diploscapter pachys*) (Figura S1).

Por otra parte se llevó a cabo la predicción de la estructura secundaria, utilizando los programas Quick2D (Figura 1) y PSIPRED, de sus respectivos servidores online comparando sus resultados. Para visualizar si la proteína posee segmentos transmembrana, se utilizó el servidor Phobius (Figura 2.a). Así mismo para conocer las posibles regiones desordenadas, se utilizó el servidor IUPred2A (Figura 2.b). La búsqueda de dominios se llevó a cabo mediante la base de datos de Pfam. La proteína en estudio presenta un dominio que abarca desde la posición 6 a la posición 88 en la secuencia. El dominio se denomina DUF5076, y en la base de datos mencionada se pueden observar las secuencias (153), las arquitecturas (3), las especies (128) y las estructuras (4) asociadas al mismo.

Como se observa en la figura 2.a, la probabilidad asignada para conocer si la proteína posee segmentos transmembrana es muy baja, y ocupa aproximadamente desde la posición 20 a 40 de la secuencia de aa. El programa Quick2D al tener integrada la predicción de segmentos transmembrana, y al ser la probabilidad

muy baja, no lo muestra. En la figura 2.b se puede observar que la secuencia presenta regiones más desordenadas desde la posición 1 a 15 y desde la 83 a la 100, aproximadamente, como también lo muestra en coincidencia el Quick2D.

Continuando con el análisis secuencial, para la realización del alineamiento múltiple, se seleccionaron 15 secuencias que se muestran en la Tabla S1. Para la realización del mismo se utilizó el programa MUSCLE, del servidor EMBL-EBI.

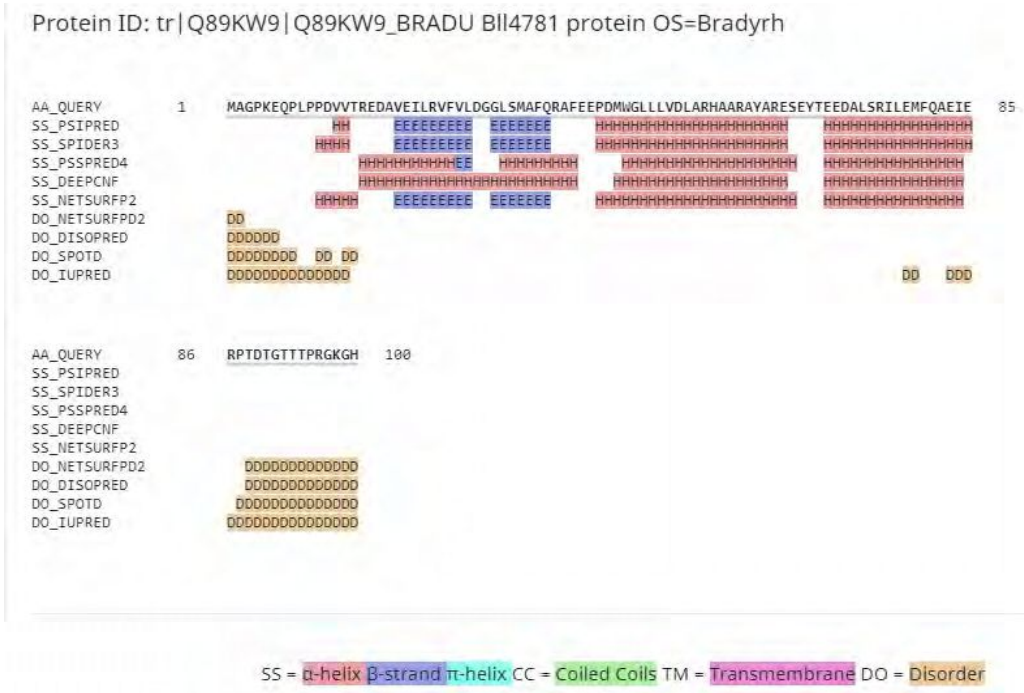


Figura 1. Predicción de estructura secundaria por Quick2D.

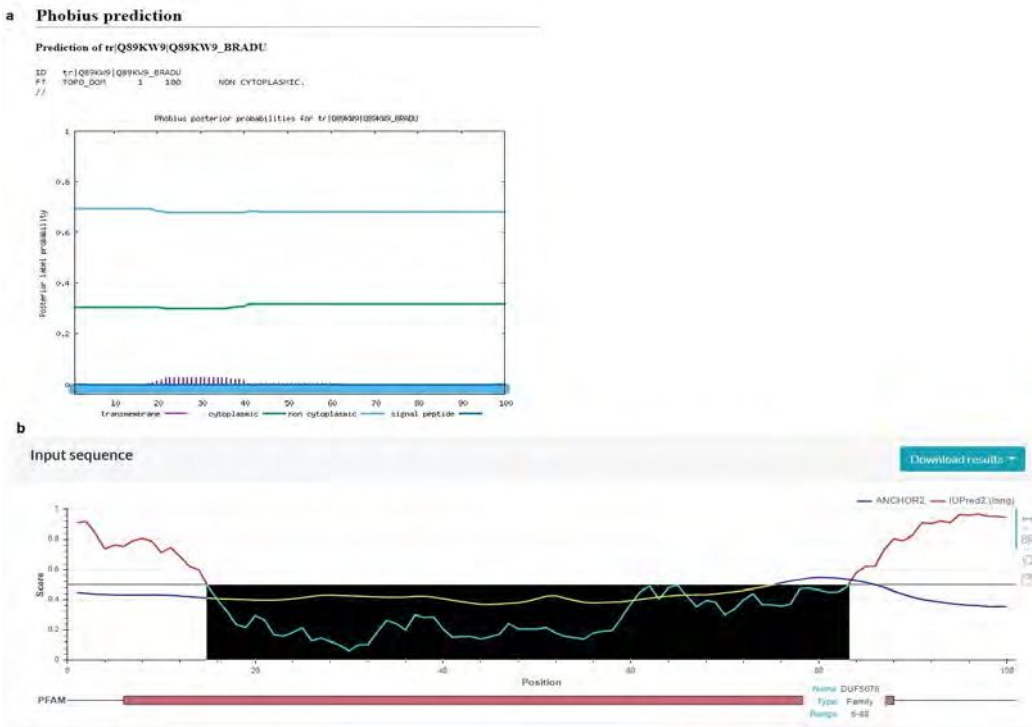


Figura 2. Predicción de estructura secundaria y regiones desordenadas. a. Predicción de segmentos transmembrana por Phobius. b. Predicción de regiones desordenadas según IUPred2A.

En la Figura 3 se puede observar el resultado del alineamiento múltiple:



Figura 3. Alineamiento múltiple en formato CLUSTAL de la secuencia *query* y las seleccionadas.

Como se observa en la figura anteriormente presentada, en todas las secuencias fue necesario la introducción de gaps para un correcto alineamiento. Por otra parte, 9 posiciones son idénticas entre todas las secuencias, mientras que 13 son más o menos similares según el programa.

Por otro lado, a través del servidor ConSurf se pudieron observar las variaciones de las posiciones, es decir, de las menos conservadas a las más conservadas. En la figura S2 se presenta una imagen con la descripción anterior.

En la etapa siguiente se obtuvo un modelo tridimensional utilizando el programa Modeller (Sali & Blundell, 1993). Para ello se buscaron secuencias homólogas a la *query sequence* que presentarán una estructura conocida utilizando los programas BLAST y PSI-BLAST, HHpred y el programa FFAS03 (Fold & Function Assignment) (Godzik Lab). De esta forma se obtuvo una secuencia *template* con 76% de identidad, un E-value de $7,9 \times 10^{-35}$ y un Score de 192,7 en el programa HHpred y 75% de identidad y un Score de -61.7 con el programa FFAS03. El *template* utilizado corresponde al ID 3IC3 del PDB (Protein Data Bank), nomenclatura que corresponde a una proteína de la bacteria fotosintética *Rhodospseudomonas palustris* CGA009, la cual interacciona con 5 ligandos (beta-D-glucopiranososa, iones fosfato, sodio y potasio, y 1,2-etanodiol). El alineamiento entre la *query sequence* y la secuencia de 3IC3 se llevó a cabo utilizando el programa T-Coffee (Figura S3).

Para obtener un modelo estructural de la secuencia en estudio, se ejecutó el programa Modeller (versión 10.1), y se escogió como valor arbitrario un número de 5 modelos a obtener y evaluar. Para la visualización de los mismos se utilizó el programa PyMol, se compararon y evaluaron los modelos por el potencial DOPE antes y después del refinamiento de los *loops* y por el servidor ProSa (Wiederstein & Sippl, 2007). En la figura S4 se muestra el diagrama para la evaluación global y local (perfil energético).

A continuación, se muestra el resultado del alineamiento entre el mejor modelo generado por Modeller y la estructura del *template* junto a los ligandos (Figura 4). Se puede observar que las estructuras tienen alta semejanza debido al alto grado de conservación de la secuencia. A su vez, para clasificar estructuralmente a la proteína se realizó una búsqueda en la base de datos CATH, obteniéndose la siguiente clasificación a nivel Superfamilia: CATH Superfamily 3.30.2370.10 (putative pyruvate dehydrogenase). Para predecir la posible función biológica de la proteína estudiada, a partir de la información secuencial y estructural, se realizó una

búsqueda exhaustiva en diferentes bases de datos para tener una aproximación a conocer la respuesta a la pregunta de ¿qué función cumple esta proteína? La búsqueda se basó en el seguimiento del nº EC (Enzyme Commission), el cual para la proteína estudiada según el servidor ExPASy es el 1.2.2, que refiere a enzimas oxidoreductasas actuantes sobre donantes de grupos aldehídos u oxo con citocromo como aceptor.

Sin embargo, en otras bases de datos como BRENDA, IUBMB Enzyme Nomenclature, IntEnz y Gene Ontology (GO), este EC se muestra como nº 1.2.5.1, con el nombre aceptado de piruvato deshidrogenasa (quinona), donde el cambio se produce en que la enzima actúa sobre donantes de grupos aldehídos u oxo con una quinona o componente similar como aceptor, participando en las rutas metabólicas de la fermentación del acetato y metabolismo del piruvato, realizando la posible función molecular de catalizar la siguiente reacción:

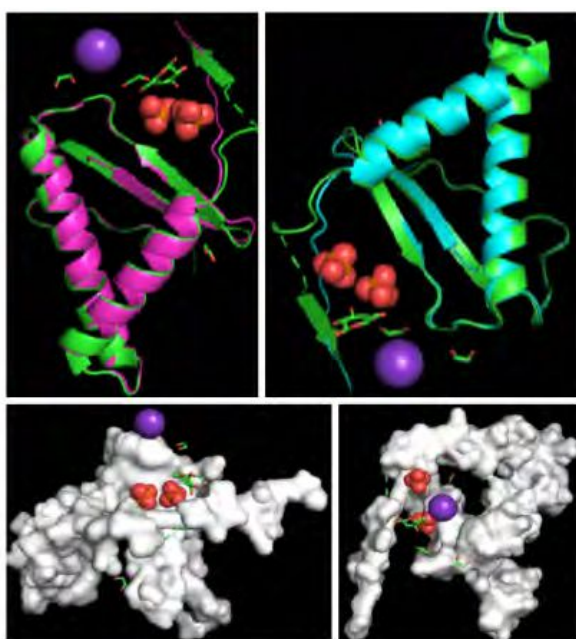
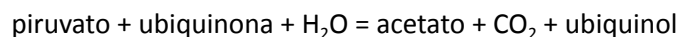


Figura 4. Modelo propuesto para la proteína bll4781 de *Bradyrhizobium diazoefficiens*. Se realizó el modelado por homología utilizando el programa Modeller y la proteína 3IC3 (PDB) como template. Las figuras se obtuvieron utilizando el programa PyMol.

Finalmente, se realizó un estudio filogenético con el objetivo de conocer las relaciones y distancias evolutivas entre las secuencias (especies) seleccionadas para tal fin. Como primer paso para la realización del árbol filogenético, se realizó un alineamiento múltiple entre la secuencia de la proteína de estudio y las secuencias homólogas seleccionadas en BLAST (16 secuencias en total). Se utilizó el servidor T-Coffee para la realización del mismo y para evaluar su calidad. En la figura S5 se observa el alineamiento múltiple obtenido en T-Coffee.

Para la construcción del árbol filogenético se hizo uso del programa PhyML 3.1, utilizando bootstrap para el soporte de los nodos (100 replicantes), y utilizando como árbol inicial un árbol construido por BioNJ (Neighbor-Joining). El mejor modelo evolutivo para la construcción del árbol fue WAG (Whelan & Goldman) según el programa ModelTest (paquete HYPHY) considerando un número de 4 categorías de velocidades de sustitución para la *gamma distribution*. En la siguiente figura (Figura 5) se observa el árbol filogenético obtenido, con los nombres de las respectivas especies, nodos, y longitud de ramas.

CONCLUSIONES Y DISCUSIÓN

La proteína bli4781 perteneciente a la bacteria de suelo Gram negativa *Bradyrhizobium diazoefficiens* presenta una longitud de 100 aa en su secuencia. Al realizar una búsqueda de homólogos, el mayor porcentaje de los mismos presentan longitudes de secuencia muy similares y el dominio DUF5076, de función desconocida. Con el programa HHpred, se encontró que el primer *hit* presentó la mayor significancia estadística y el porcentaje de identidad más alto (%76). Esta búsqueda se realizó contra la base de datos del PDB y la secuencia y estructura asociada demuestra que pertenece a la proteína homotetrámero piruvato dehidrogenasa de la bacteria fotosintética *Rhodopseudomonas palustris* CGA009. A partir del alineamiento realizado entre el *target* y la secuencia del *template* se puede observar un alto grado de conservación, lo que se demuestra en las imágenes del modelado por homología realizado en el programa Modeller.

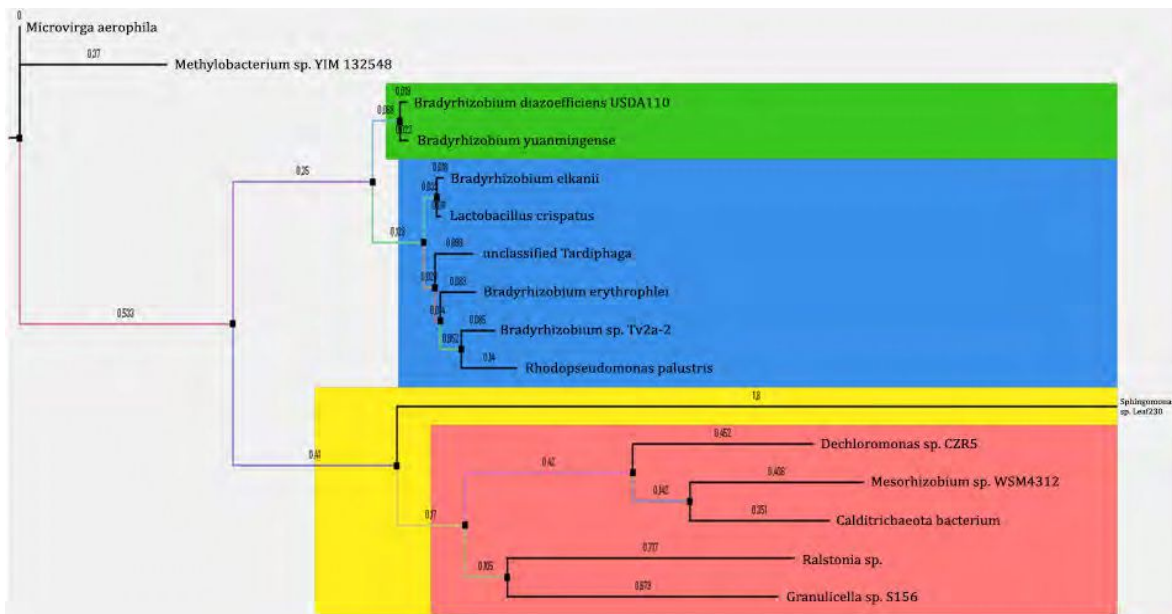


Figura 5. Árbol filogenético de 16 secuencias homólogas a la proteína bli4781. El árbol fue obtenido utilizando PhyML 3.1 y visualizado en FigTree.

La mayor porción de la proteína no presenta regiones desordenadas, las mismas se observan desde las posiciones 1 a 15 y 83 a 100, respectivamente. Así mismo, esta proteína carece de regiones transmembrana.

En las relaciones filogenéticas, se observa que *Bradyrhizobium diazoefficiens* comparte estrecha relación con *Bradyrhizobium yuanmingense* ya que divergieron en conjunto a partir de su ancestro común, formando parte de un clado más divergente y diverso junto a *Bradyrhizobium elkanii*, *Bradyrhizobium erythrophlei*, *Bradyrhizobium sp. Tv2a-2*, *Lactobacillus crispatus*, *Tardiphaga* y *Rhodopseudomonas palustris*.

Ante la ausencia de datos que puedan confirmar la función de la proteína estudiada, se le asignó la posible función molecular de catalizar la reacción química entre el piruvato y ubiquinona dando como producto final acetato, ubiquinol y dióxido de carbono.

BIBLIOGRAFÍA

A. Šali & T. L. Blundell. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234, 779-815.

[BRENDA, the ELIXIR core data resource in 2021: new developments and updates.](#) Chang A., Jeske L., Ulbrich S., Hofmann J., Koblitz J., Schomburg I., Neumann-Schaal M., Jahn D., Schomburg D., *Nucleic Acids Res.*, 49:D498-D508

- Brooksbank, C., Camon, E., Harris, M. A., Magrane, M., Martin, M. J., Mulder, N., ... & Cameron, G. 2003. The European Bioinformatics Institute's data resources. *Nucleic acids research*, 31(1), 43-50.
- CATH: increased structural coverage of functional space. Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, Pang CSM, Woodridge L, Rauer C, Sen N, Abbasian M, Le Cornu S, Lam SD, Berka K, Varekova IH, Svobodova R, Lees J, Orengo CA. *Nucleic Acids Res.* 2021. [Pubmed: 33237325](#). [doi: 10.1093/nar/gkaa1079](#)
- Davis-Richardson, A. G., Russell, J. T., Dias, R., McKinlay, A. J., Canepa, R., Fagen, J. R., ... & Triplett, E. W. 2016. Integrating DNA methylation and gene expression data in the development of the soybean-Bradyrhizobium N₂-fixing symbiosis. *Frontiers in microbiology*, 7, 518.
- Gábor Erdős & Zsuzsanna Dosztányi. 2020. [Analyzing Protein Disorder with IUPred2A](#). *Current Protocols in Bioinformatics* 2020; 70(1):e99
- Gasteiger E., Gattiker A., Hoogland C., Ivanyi I., Appel R.D., Bairoch A. 2003. ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* 31:3784-3788
- H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Research*, 28: 235-242. Disponible en: <https://www.rcsb.org/>
- Jaroszewski, L., Rychlewski, L., Li, Z., Li, W. & Godzik, A. 2005. FFAS03: a server for profile-profile sequence alignments. *Nucl. Acids Res.* 33, W284-W288
- Kaneko, T., Nakamura, Y., Sato, S., Minamisawa, K., Uchiumi, T., Sasamoto, S., ... & Tabata, S. 2002. Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA110. *DNA research*, 9(6), 189-197
- Landau M., Mayrose I., Rosenberg Y., Glaser F., Martz E., Pupko T. and Ben-Tal N. 2005. ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucl. Acids Res.* 33:W299-W302.
- National Center for Biotechnology Information (NCBI). Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. 2021. <https://www.ncbi.nlm.nih.gov/>
- Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). 2021. School of Biological and Chemical Sciences, Queen Mary, University of London. <https://www.qmul.ac.uk/sbcs/iubmb/>
- [Pfam: The protein families database in 2021](#): J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G.A. Salazar, E.L.L. Sonnhammer, S.C.E. Tosatto, L. Paladin, S. Raj, L.J. Richardson, R.D. Finn, A. Bateman. 2020. *Nucleic Acids Research*. doi: 10.1093/nar/gkaa913
- Phobius: A combined transmembrane topology and signal peptide predictor. 2021. Centre Stockholm Bioinformatics. Disponible en: <https://phobius.sbc.su.se/>
- Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. Gabler F, Nam SZ, Till S, Mirdita M, Steinegger M, Söding J, Lupas AN, Alva V. [Curr Protoc Bioinformatics. 2020 Dec;72\(1\)](#)
- [The InterPro protein families and domains database: 20 years on](#). Matthias Blum, Hsin-Yu Chang, Sara Chuguransky, Tiago Grego, Swaathi Kandasamy, Alex Mitchell, Gift Nuka, Typhaine Paysan-Lafosse, Matloob Qureshi, Shriya Raj, Lorna Richardson, Gustavo A Salazar, Lowri Williams, Peer Bork, Alan Bridge, Julian Gough, Daniel H Haft, Ivica Letunic, Aron Marchler-Bauer, Huaiyu Mi, Darren A Natale, Marco Necci, Christine A Orengo, Arun P Pandurangan, Catherine Rivoire, Christian J A Sigrist, Ian Sillitoe, Narmada Thanki, Paul D Thomas, Silvio C E Tosatto, Cathy H Wu, Alex Bateman, Robert D Finn *Nucleic Acids Research*. 2020, gkaa977, PMID: [33156333](#)
- The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. <https://www.uniprot.org/>
- Wiederstein, M. & Sippl, M. J. 2007. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic acids research*, 35(suppl_2), W407-W410.

Implicancias del receptor ionotrópico NMDA subunidad 3A en la esquizofrenia

Iris Quimey López

Cátedra de Bioinformática, Área de Biotecnología y Biología Molecular, Departamento de Ciencias Biológicas, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, La Plata, Argentina.

RESUMEN

El rápido avance de la bioinformática nos da nuevas formas de analizar enfermedades y trastornos para los que aún no se han podido dilucidar tanto sus causales, como nuevas estrategias de tratamiento más efectivas. Las herramientas bioinformáticas permiten un análisis profundo de las proteínas a nivel secuencial y estructural. Con esta metodología se ha analizado la subunidad proteica 3A del receptor ionotrópico NMDA para encontrar posible información sustancial acerca de su funcionamiento, y variabilidad genética que pueda estar implicada en el desarrollo de enfermedades y trastornos neurodegenerativos, particularmente en la esquizofrenia.

Este informe analiza de forma exhaustiva la secuencia de la proteína, su estructura, su desarrollo evolutivo, y sus posibles funciones biológicas, encontrándose una estructura altamente conservada, con un segmento C-terminal con un alto grado de desorden convirtiéndose en una fuente de variabilidad genética importante, propensa a aumentar la frecuencia de desarrollo de enfermedades neurodegenerativas, y posiblemente implicada en la evolución del cerebro social humano por su papel en el refinamiento sináptico. En cuanto a su función molecular se predice que es un componente intrínseco de la membrana, por ser un receptor de NMDA de canales iónicos activados por glutamato. La proteína consta de al menos cuatro dominios: la familia Lig_chan (receptor ionotrópico de glutamato), la familia Lig_chan-Glu_bd (región con canal iónico ligado L-glutamato, y el sitio de unión a glicina), la familia ANF_receptor (región de unión del ligando de la familia de receptores), y el dominio de proteínas de unión periplásmica tipo 1 y 2. Además se incluye junto al análisis estructural, un modelado de la proteína Q8TCU5 utilizando como template la proteína 7KS0.

PALABRAS CLAVE: GRIN 3A; receptor NMDA; esquizofrenia; primates ; bioinformática.

INTRODUCCIÓN

En la actualidad disponemos de herramientas bioinformáticas que nos permiten averiguar datos de interés sobre una secuencia determinada, predecir su estructura y lograr una aproximación a la predicción de su función biológica. Esto se logra por medio de la aplicación de algoritmos destinados a producir alineamientos por similitud (secuencial y/o estructural) entre las secuencias proteicas acumuladas en la base de datos elegida por el investigador, o bien una combinación de múltiples bases de datos, y analizando estos resultados con criterio biológico. La precisión del alineamiento es resultado directo del algoritmo utilizado (según la matriz en que se base el método, con sus respectivos parámetros), y de la cantidad y calidad de las secuencias de proteínas homólogas utilizadas. En este trabajo utilizaremos diversos métodos bioinformáticos para la caracterización del receptor ionotrópico de glutamato NMDA subunidad 3A de homo sapiens (UniprotID: Q8TCU5). A continuación se muestra un modelo esquemático del complejo proteico NMDA.

Hipótesis sobre la esquizofrenia

La esquizofrenia es un trastorno mental caracterizado por presentar síntomas clasificados como: de primer rango, que incluyen pensamiento sonoro, voces que discuten, experiencia de pasividad somática, influencia, imposición y robo del pensamiento, transmisión de pensamiento, percepciones delirantes, cualquier experiencia que implique voluntad, afectos e impulsos dirigidos; y de segundo rango, que incluyen otros trastornos de la percepción, ideas delirantes súbitas, perplejidad, cambios depresivos o eufóricos, sentimientos de empobrecimiento emocional.

El mecanismo neuroquímico subyacente a la esquizofrenia aún permanece desconocido, pero se considera un trastorno altamente heredable gracias a diversos estudios realizados en familiares cercanos a pacientes que lo padecen. Se han realizado exhaustivas búsquedas de alelos relacionados (preferentemente en genes involucrados en la sinapsis), por ejemplo, se han encontrado polimorfismos de nucleótido único (SNP) en un gen que codifica el receptor de dopamina D2, DRD2 (objetivo de los fármacos antipsicóticos) y muchos genes implicados en las vías de los neurotransmisores de glutamina y la plasticidad sináptica (p. Ej., GRM3, GRIN2A, SRR, GRIA1). En otro estudio se utilizó una matriz personalizada de 1536 SNPs para interrogar a 94 genes candidatos funcionalmente relevantes para la esquizofrenia e identificar asociaciones con 12 endofenotipos neurofisiológicos y neurocognitivos hereditarios, finalmente se observaron asociaciones con endofenotipos para 46 genes de potencial significado funcional, con tres SNPs. Los resultados apoyan colectivamente un papel importante de los genes relacionados con la señalización del glutamato en la mediación de la susceptibilidad a la esquizofrenia (Greenwood Tiffany A. et. al 2011).

La dopamina fue el primer objetivo de los mecanismos de acción de los fármacos antipsicóticos y la capacidad psicotrópica de las sustancias psicoestimulantes desde que se desarrolló la hipótesis dopaminérgica. Actualmente la hipótesis dopaminérgica junto a la hipótesis glutaminérgica son consideradas las más importantes y de mayor base empírica para el estudio de la esquizofrenia y sus espectros (Gurillo Muñoz P. 2016).

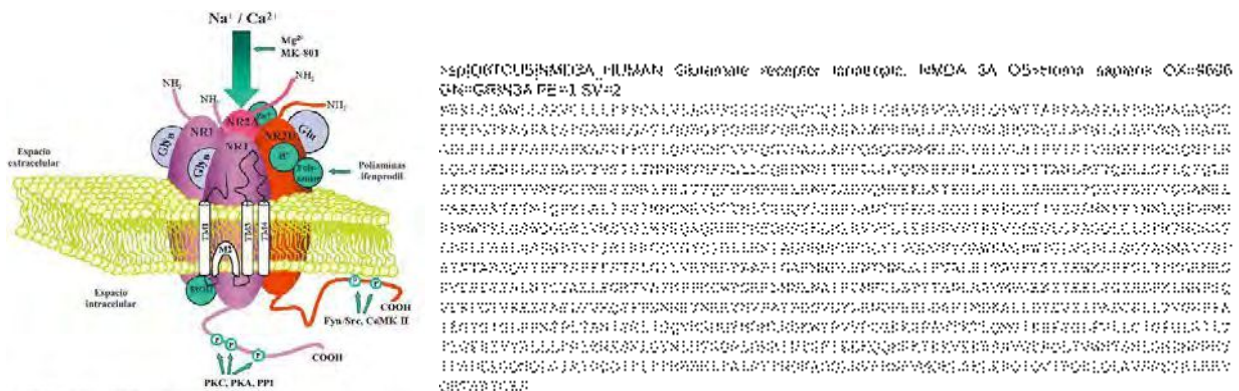


Figura 1 (a). Modelo esquemático del receptor NMDA, y figura 1-B secuencia de la subunidad 3A del receptor NMDA. A la izquierda se representa el ensamblaje heteromérico de 4 subunidades en los receptores NMDA. Cada subunidad tiene cuatro regiones hidrofóbicas, y tres formas de dominio transmembrana (TM1, TM3, TM4). M2 toma la conformación de horquilla dentro de la membrana y forma el poro. Los complejos funcionales del receptor están formados por combinaciones de las subunidades NR1 y NR2, las cuales contienen los sitios de unión de los agonistas, algunos antagonistas y otros moduladores. Fuente: Jozsef Nagy, tipo de subunidades NR2B del receptor NMDA; un blanco posible para el tratamiento de la dependencia del alcohol. Current Drug Targets CNS Neurol Disord., 2004,3(3):169-79. (b) A la derecha se exhibe la secuencia en formato fasta.

En relación a la hipótesis glutaminérgica diversos servidores aportan estudios basados en la búsqueda de genes (OMIM), y estos relacionan el desarrollo de la esquizofrenia con la participación del gen GRIN3A, en

conjunto con los genes señalados anteriormente, en el desarrollo de los trastornos esquizoafectivos (esquizofrenia en combinación con el trastorno bipolar, y la depresión).

Existen diversas hipótesis acerca del desarrollo de este trastorno, entre ellas la de la insuficiente 'poda sináptica', o bien el exceso de poda neuronal (De las Matas Martín, y Del Carmen María, 2014).

Entre las hipótesis evolutivas podemos encontrar aquella que relaciona la evolución del cerebro social humano con la esquizofrenia:

La esquizofrenia puede ser una compensación de la evolución de la inteligencia social, la cual conlleva un dramático incremento en la conectividad cortical de los primates. Se sugiere un modelo de dos pasos en el trasfondo evolutivo de la esquizofrenia. El primer paso evolutivo se dio hace unos 5 ó 6 millones de años, hacia una más compleja conectividad inter- e intrahemisférica. Un segundo paso fue hace, aproximadamente, 150.000 años, cuando alguna desconocida mutación incrementó la vulnerabilidad de tales conexiones, lo que podría estar asociado con la evolución de la metacognición y la 'teoría de la mente'. De acuerdo con el autor Burns, la esquizofrenia podría ser un costo de la evolución del cerebro social humano, y, en este sentido, su hipótesis se hallaría entre las que suponen que el trastorno significa una desventaja compensatoria en la evolución del cerebro social. Otro autor propone que un gen que regula la dominancia cerebral está involucrado en el origen de los trastornos psicóticos, y que en este origen podría ser trascendente en la evolución del lenguaje (Altschul, S.F, et. al, 1990).

En este trabajo se realizará una búsqueda de los homólogos tanto cercanos como remotos, con su respectiva distribución taxonómica. Con esta información se pretende predecir la estructura secundaria, los segmentos transmembrana, regiones desordenadas, dominios, además de proporcionar un modelo estructural de la proteína, construir un árbol filogenético, y de ser posible predecir la función biológica de la proteína. Con este análisis se busca enriquecer el conocimiento que se tiene de esta proteína en relación a su función neurotransmisora y los trastornos que están arraigados a su alteración, así como la función biológica de sus homólogos más cercanos, y de ser posible detectar las diferencias secuenciales y/o estructurales que causen la ausencia de estos trastornos en el resto de los primates (en base a las teorías evolutivas de la esquizofrenia y la evolución del cerebro social humano, anteriormente citadas), por lo cual se le considera un trastorno propio de la especie humana.

MÉTODOS Y RESULTADOS

Búsqueda de secuencias homólogas

Para comenzar con la caracterización del receptor ionotrópico de glutamato NMDA subunidad 3A, se realizó la búsqueda de homólogos por Blast (Altschul S.F., 1997). Blast es una herramienta básica de búsqueda de secuencias similares basada en alineación local. El programa compara una secuencia (ya sea nucleotídica o proteica) contra las secuencias contenidas en una base de datos seleccionada (método de comparación secuencia a secuencia), calculando a su vez la significancia estadística (e-value) para cada resultado. Este algoritmo prioriza la velocidad en vez de la sensibilidad de la búsqueda, y se basa en la comparación de K-Tuples. Al realizar la búsqueda con los parámetros por defecto se obtienen 5074 resultados de posibles homólogos. Siendo Blast un algoritmo de alineamiento local, encontrará homólogos tanto de la proteína completa como de sus dominios. Para quedarnos solo con homólogos correspondientes a la proteína completa utilizamos los siguientes filtros y valores para los parámetros: Coverage > 70%, %Identity > 40%, Expected threshold = 100, Matriz: compositional score matrix adjustment, Database: All non-redundant GenBank CDS translations+PDB+SwissProt +PIR+PRF excluding environmental samples from WGS projects.

De esta forma se encontraron 1254 posibles homólogos. Considerando los resultados del alineamiento obtenidos con un query coverage (>70%) y con un cutoff correspondiente a un %ID>30 los alineamientos encontrados son homólogos por tener una distancia evolutiva cercana que puede deducirse con el análisis secuencial, sin necesidad de realizar un análisis estructural.

Adicionalmente, Blast aporta toda la información taxonómica de las proteínas alineadas, el dominio, el género, la especie, orden, etc. Con el análisis taxonómico podemos observar que los hits con mayor significancia estadística, es decir, con mayor score (hasta 2315 score total) pertenecen a la especie *Homo sapiens*, los 47 homólogos más relevantes que le siguen son también pertenecientes a la orden de los primates. Los hits de menor score pertenecen mayormente a diversos vertebrados, especialmente mamíferos. Se obtuvo como homólogo más lejano al receptor ionotrópico del glutamato, NMDA 3A del organismo *Crassostrea virginica* (%ID de 25,96, e-value=1e-82, y query coverage del 74%).

Con el objetivo de buscar homólogos remotos se realizaron búsquedas utilizando PSI-BLAST (Profile – secuencia) y HMMER (HMM) PSI-BLAST (o Position Specific Iterated BLAST) es un programa muy rápido que realiza un simple BLAST con una secuencia y, a partir de los resultados, construye un perfil o PSSM. Entonces, la siguiente búsqueda la realiza con ese perfil, lo que permitirá encontrar, idealmente, homólogos remotos. Dados los nuevos homólogos se genera un nuevo perfil, que idealmente contendrá mayor cantidad de información y podrá realizar otra búsqueda. Es un proceso iterativo. Con una búsqueda refinada (con parámetro word size=2) se obtuvieron homólogos remotos, donde el resultado con mayor distancia evolutiva encontrado fue el receptor ionotrópico 6 perteneciente al organismo *Diaphorina citri* (%ID de 23,69, e-value=1e-59, y query coverage de 75%).

HMMER proporciona herramientas para crear modelos probabilísticos de familias de dominios de secuencias de proteínas y ADN (HMM) y para usar estos perfiles para anotar nuevas secuencias, buscar en las bases de datos de secuencias de homólogos adicionales y crear alineamientos profundos de múltiples secuencias. Como resultado se encontraron 12.052 posibles homólogos (base de datos: uniprot refprot v.2019_09). Se logró llegar hasta un 25.6% de ID, con un 51.2% de similitud, y con un query coverage en el intervalo de 671-756 (Figura S1).

Predicción de estructura secundaria

Se realizó la predicción de estructura secundaria con diversos programas, incluyendo Quick2D (Gabler F, et al, [2020](#)), PSIPRED (Buchan DWA, & Jones DT, 2019), y Porter (M.Torrise, 2019). Por medio de Quick2D (Figura S2) se detectó un péptido señal potencial en su extremo N-terminal, además se predijo que la estructura secundaria estaría compuesta de manera uniforme por segmentos de hélice alfa y hoja beta plegada, y se predijeron cuatro regiones transmembrana (tres de ellas en una región intermedia de la secuencia, y la última hacia el extremo C-terminal). Además se detectaron pequeños segmentos desordenados dentro de la secuencia. Los resultados de PSIPRED, un método de predicción de estructura secundaria que incorpora dos redes neuronales de retroalimentación que realizan un análisis de la salida en función de las matrices de puntuación específicas de posición generadas por PSI-BLAST (Position Specific Iterated – BLAST) se muestran en la figura 4, S3, S4, y S5.

Finalmente, las predicciones obtenidas mediante el programa Porter 5 (M.Torrise, 2019), compuesto por un conjunto de redes neuronales recurrentes bidireccionales en cascada y redes neuronales convolucionales, concuerdan en gran proporción con la predicción de la estructura secundaria aportada por PSIPRED (Figura S6).

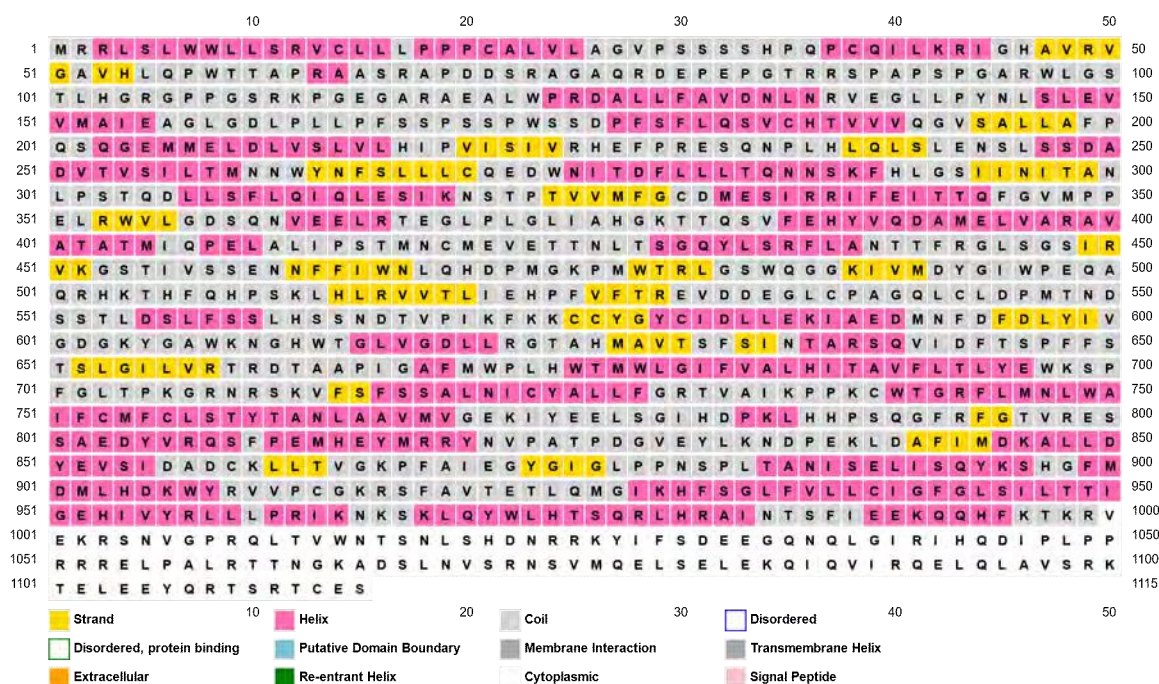


Figura 4. Predicción de estructura secundaria con PSIPRED. La figura representa la estructura secundaria predicha en cada región de la proteína. Se predijo que la proteína presenta una conformación mayormente del tipo coil, con una distribución uniforme de conformaciones tanto hélice alfa como hoja beta plegada, así como también se predice una región citoplasmática hacia el final de la secuencia. No se predijeron péptidos señal.

Regiones de baja complejidad

Las regiones de baja complejidad (Low complexity regions, LCR), son secuencias de proteínas cuya composición de aminoácidos es muy simple. Su expansión descontrolada provoca varias enfermedades humanas, incluyendo la enfermedad de Huntington y otras enfermedades neurodegenerativas y de desarrollo. Sin embargo, son sorprendentemente abundantes en las proteínas, lo que parece paradójico dado su alto potencial patógeno. Por otra parte, los datos experimentales han demostrado que la formación de nuevas LCR, o la modificación de las existentes, puede tener consecuencias funcionales. Pueden ser una importante fuente de variabilidad genética y podrían estar implicadas en los procesos de adaptación, además pueden estar involucradas en la diversificación de la proteína, ya sea proporcionando nuevas secuencias funcionales que modificarán las proteínas existentes o estando involucradas en la formación de nuevas secuencias codificantes en la proteína. Para el estudio de este tipo de secuencias se utilizó SEG: Prediction of Low Complexity Regions (Zhang M, et. al, 2018). Los resultados muestran una gran cantidad de regiones de baja complejidad en el fragmento N-inicial y el C-terminal de la secuencia de la proteína Q8TCU5 (Tabla 1), lo que implica que estas dos regiones son fuentes de alta variabilidad genética, propensas a aumentar la frecuencia de desarrollo de enfermedades neurodegenerativas.

Tabla 1. Resultados obtenidos mediante SEG.

Predicción	SEG 12 2.2 2.5	SEG 25 3.0 3.3	SEG 45 3.4 3.75
Regiones de baja complejidad	1-1; 2-26; 27-26; 61-74; 75-156; 157-181;182-547;1047-1059; 1060-1115	1-2; 3-33; 34-56; 129-156; 157-181; 182-547	1-1;2-255;256-1115

Nota: Parámetros por default para la secuencia proteica Q8TCU5, NMD3A_HUMAN Glutamate receptor ionotropic, NMDA 3A, Homo sapiens (GRIN3A).

Predicción de dominios globulares

Las proteínas globulares son proteínas formadas únicamente por aminoácidos, suelen estar compuestas de una sola molécula proteica, o unas pocas moléculas combinadas que se pliegan en forma esférica y forman una estructura más compleja. Se caracterizan por doblar sus cadenas en forma esférica compacta dejando grupos hidrófobos en el core de la proteína y grupos hidrófilos hacia afuera, lo que hace que sean solubles en disolventes polares como el agua. Forman suspensiones coloidales. La mayoría de las enzimas, anticuerpos, algunas hormonas y proteínas de transporte son globulares. Para la predicción de regiones globulares se utilizó GlobPlot (Linding R, et. al, 2003), un servicio web que permite al usuario trazar la tendencia dentro de la proteína para el orden / globularidad y el desorden. Adicionalmente identifica segmentos entre dominios que contienen motivos lineales y también regiones aparentemente ordenadas que no contienen ningún dominio reconocido. Como resultado (Figura S7), se predijeron regiones globulares en los segmentos: 240-596, 626-930, 946-1174.

Predicción de segmentos transmembrana

Para la predicción de segmentos transmembrana se utilizaron los servidores TMHMM (DTU Health Tech, 2017), TMPred (Hoffman & Stoffel, 1993) y Phobius (Lukas Käll, 2007). TMHMM se encarga de predecir las hélices transmembranas en las proteínas, está basado en un modelo oculto de Markov (HMM), donde una de las principales ventajas es que es posible modelar la longitud de la hélice estableciendo límites superiores e inferiores para la longitud de una hélice de membrana. Los HMM son muy adecuados para la predicción de hélices transmembrana porque pueden incorporar hidrofobicidad, sesgo de carga, y longitudes de hélice. Los resultados obtenidos con TMHMM se muestran en la figura 9.

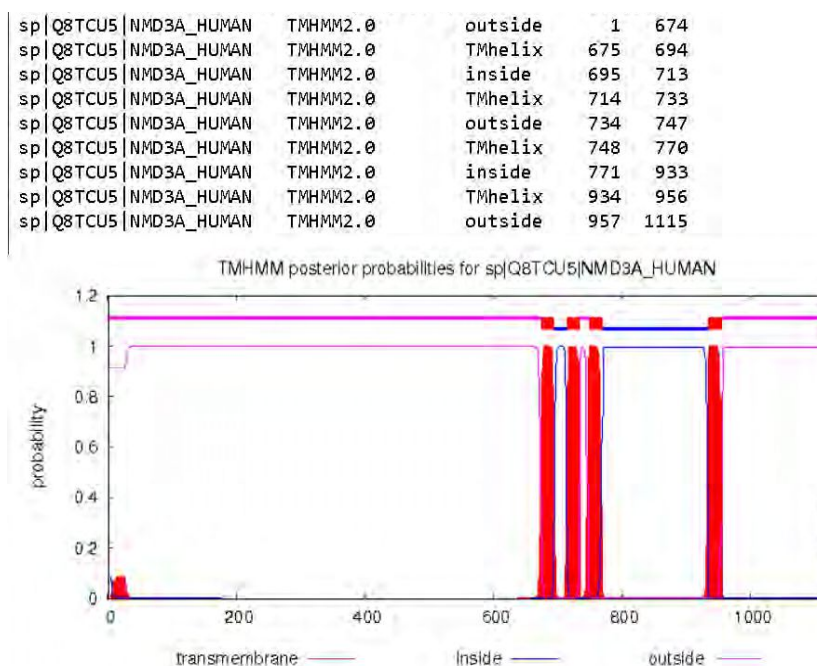


Figura 9. Probabilidad de existencia de hélices transmembrana en la secuencia proteica de Q8TCU5. La leyenda superior al gráfico indica las predicciones de segmentos transmembrana/ citoplasmáticos/ no citoplasmáticos para cada segmento de la proteína, luego el gráfico inferior representa la probabilidad de encontrar regiones transmembrana en determinados segmentos de la secuencia, evidenciándose una mayor probabilidad de encontrar regiones transmembrana en los segmentos 675-694, 714-733, 748-770, y 934-956, lo que podría ser una hélice alfa múltiple. Hay una alta probabilidad de que esta proteína contenga 4 regiones transmembrana.

TMPred es una base de datos de proteínas transmembrana y dominios helicoidales que atraviesan la membrana. TMPred originalmente era una herramienta para analizar las propiedades de las proteínas transmembrana. Se basa principalmente en SwissProt, pero también contiene información de otras bases de datos. TMPred se utilizó con los siguientes parámetros: matriz=MTIDK; Window width: 14,21, 28; Ponderación de las posiciones= no. Este programa indica que hay dos modelos posibles basados en las predicciones de segmentos transmembrana (Figura S8): Modelo 1, con 7 hélices transmembranas (Score=10975); y el modelo 2, con 6 hélices transmembranas (Score=9902).

Adicionalmente se utilizó el servidor Phobius (Figura 10), que sirve para predecir la topología transmembrana y los péptidos señal de la secuencia de aminoácidos de una proteína.

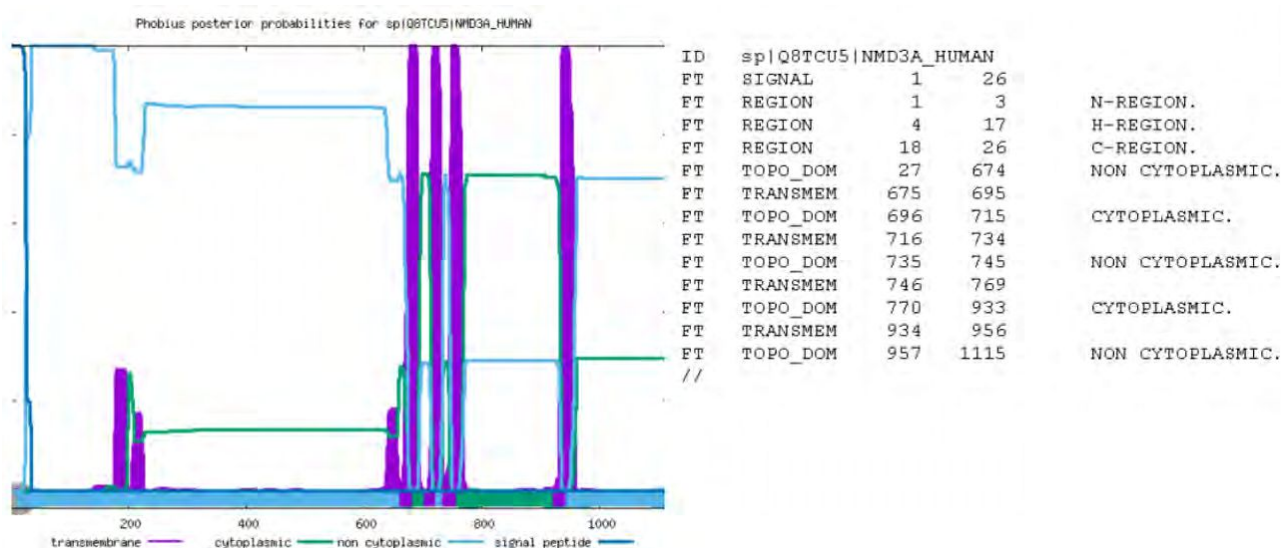


Figura 10. Predicción de la topología transmembranal. El gráfico se generó calculando la probabilidad total de que un residuo pertenezca a una hélice citoplasmática o no citoplasmática, hélice TM, o péptido señal, sumada en todos los caminos posibles a través del modelo, utilizándose parámetros por default. Aquí uno puede ver posibles hélices TM débiles que no fueron predichas. En la figura se observan al menos 4 segmentos coiled coil hacia el final de la cadena, entre las posiciones que van del 600 a la 800 y de la posición 1000 a la 1115. Esto indica que hay una región citoplasmática en la misma región donde se sugería la existencia de hélices alfa múltiples en TMHMM, es decir, en el segmento desde la posición 675 hasta la 735 y el segmento que va de 770 a la posición 956. Además me indica la posible existencia de una pequeña hélice alfa en el segmento que va desde la posición 4 a la 17. Como conclusión, podemos decir que existen al menos 4 regiones transmembranales (que coinciden con las regiones globulares predichas), y un péptido señal en la región N-inicial (denotada en color azul en la figura).

Predicción de regiones coiled-coil

Para la predicción de segmentos con estructura de coiled-coil se utilizaron los servidores COILS (Lupas, 1991) (Figura 11), Paircoil (Bonnie Berger, 1995) (Figura 12) y Paircoil2 (McDonnell, 2006). COILS es un programa que compara una secuencia con una base de datos de coiled-coils de dos hebras paralelas conocidas y obtiene un score de similitud. Al comparar esta puntuación con la distribución de scores en proteínas globulares y en espiral, el programa calcula la probabilidad de que la secuencia adopte una conformación de hélice.

El programa Paircoil toma tres argumentos: un nombre para la secuencia (opcionalmente), un límite de probabilidad y la secuencia de aminoácidos. El límite de probabilidad determina qué tan estrictamente el programa filtra la secuencia de entrada al detectar la existencia de un dominio de coiled-coil. Se determinó empíricamente, que el valor predeterminado de 0,5 para el límite de probabilidad funciona bien. Por último

se utilizó Paircoil2, que predice el pliegue de coiled-coils a partir de la secuencia utilizando probabilidades de residuos por pares con el algoritmo [Paircoil](#) y una base de datos de coiled-coil actualizada.

Por medio del análisis predictivo dado por estos servidores a partir de la secuencia de la proteína Q8TCU5, podemos decir que hay al menos 4 regiones transmembrana en los segmentos 675 al 694, 714-733, 748-770, y el segmento final es una región coiled-coil desde la posición 1000 a la 1115.

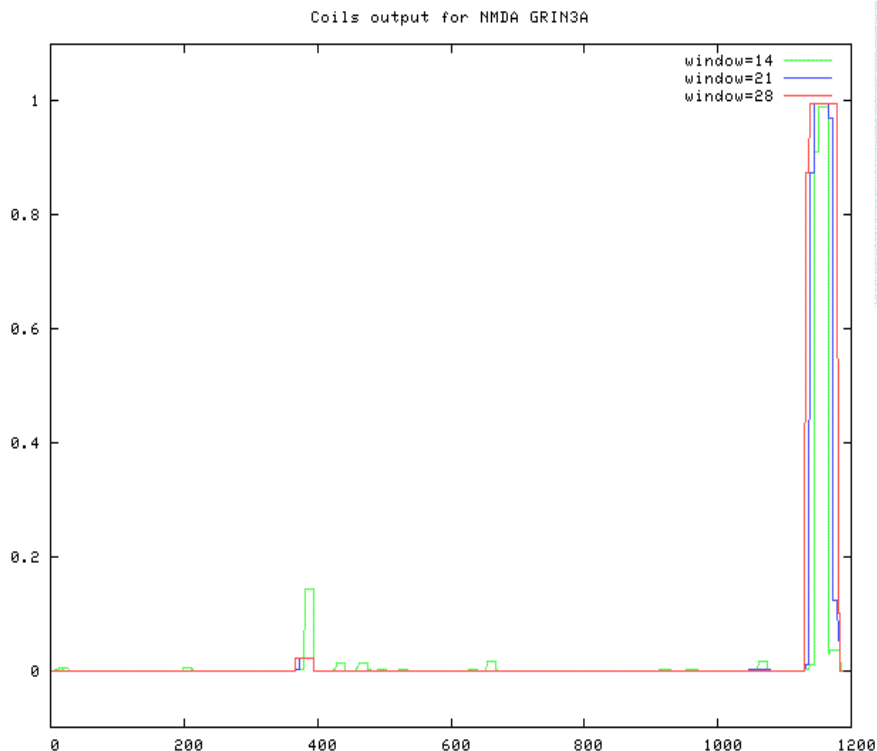


Figura 11. Probabilidad de encontrar una hélice superenrollada, y la probabilidad de formar espirales dobles en la secuencia. En el eje vertical están representadas las probabilidades (scoreadas por MTK, MTIDK, MTK_W y MTIDK_W en ese orden), en el eje horizontal el largo de mi secuencia. Los valores se obtienen con una ventana de escaneo de 21 residuos que detecta los extremos de los segmentos de espirales dobles con más precisión que la ventana de 28. Se observa la presencia de una hélice superenrollada en el segmento final de mi secuencia con un alto nivel de probabilidad. Se utilizaron los parámetros por default: Matriz=MTIDK Ponderación de las posiciones=no NCOILS versión .0 [ISREC-Server].

Predicción de regiones desordenadas

Para la predicción de regiones desordenadas se utilizaron IUPred2A (Bálint Mészáros, 2018) (Figura 13), DynaMine (Figura S9), y MobiDB (Figura S10). IUPred2A es una interfaz web combinada que permite identificar regiones de proteínas desordenadas (no tienen una estructura terciaria bien definida en condiciones nativas) usando IUPred2 y regiones de unión desordenadas usando ANCHOR2.

DynaMine (Elisa Cilia, et. al, 2013) es un predictor de la dinámica de la columna vertebral de las proteínas. Utiliza información secuencial proteica como entrada con un gran potencial para distinguir regiones de diferente organización estructural, como dominios plegados, enlazadores desordenados, glóbulos fundidos y motivos de unión preestructurados de diferentes tamaños. MobiDB (Piovesan D, et. al, 2020) proporciona información sobre regiones intrínsecamente desordenadas (IDR) y características relacionadas de varias fuentes y herramientas de predicción. Los diferentes niveles de confiabilidad y las diferentes características se informan como anotaciones diferentes e independientes.

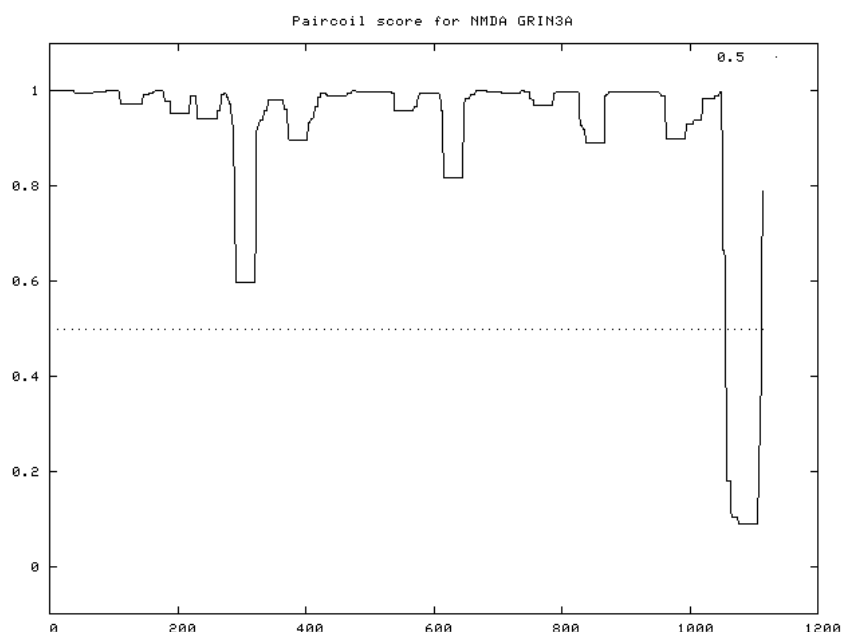


Figura 12. Probabilidad residuo por residuo. Representa la probabilidad residuo por residuo de que ese residuo se encuentre en una coiled-coil. El eje x contiene la ubicación del residuo (desde el principio, siendo 1 el primer residuo en la secuencia), y el eje y contiene la probabilidad de una bobina enrollada. El límite de probabilidad se muestra como una línea discontinua. Este gráfico indica la alta probabilidad de encontrar un segmento coiled coil en la posición 1000 a la 1115. En el resto de la secuencia se observan otros picos de menor probabilidad distribuidos en las regiones que otros programas señalaron como transmembranas. Se utilizaron los parámetros por default.

El análisis a través de los servidores citados (dynamine, MobiDB, IUPred2A) arroja resultados de gran similitud donde sugieren que mi proteína tiene un bajo porcentaje de desorden, excepto por el segmento inicial de mi proteína, de posiciones 50-150 aminoácidos (que es la región que presenta mayor desorden) y el segmento final de 1000-1115 aminoácidos. El resto de la secuencia parece tener una estructura muy conservada y por ende predecible, lo que indica que no guarda grandes distancias evolutivas con sus homólogos. Como conclusión, podemos destacar la presencia de una región de unión desordenada en el segmento que va de la posición 1 a la 200 aproximadamente, región donde se ubicaba el péptido señal anteriormente predicho. Y además se identificó a la región c-terminal como desordenada, lo que convierte a este segmento en una fuente de variabilidad genética importante.

Búsqueda de regiones repetitivas

En cuanto al análisis de la secuencia en la base de datos de proteínas repetitivas en tándem anotadas, RepeatsDB (Paladin L., et. al, 2021) proporciona la posición de la unidad, clasificación y referencia a otras bases de datos) arrojó cero resultados en su búsqueda, lo que sugeriría que esta proteína no posee una unidad repetitiva en su estructura que permita clasificarla como una proteína repetitiva.

Búsqueda de dominios conservados y motivos secuenciales

Se realizó la búsqueda de dominios conservados mediante PFAM (Mistry, J, et. al, 2021), InterPro (Blum et. al, 2020), CCD (Lu S, et. al 2020), CDART (Geer LY, 2002) y Prosite (Sigrist, et. al, 2012). Pfam es una colección de alineamientos secuenciales múltiples y de perfiles de modelos ocultos de Markov (HMM). Cada Pfam HMM representa una familia o dominio de proteínas. Al buscar una secuencia de proteínas en la biblioteca Pfam de HMM, puede determinar qué dominios lleva, es decir, su arquitectura de dominio. Pfam también se puede utilizar para analizar proteomas y arquitecturas de dominio más complejas. Como resultado de la

búsqueda se encontraron 3 dominios (Figura S11): Lig_chan (PF00060, Receptor ionotrópico de glutamato, 674-942); Lig_chan-Glu_bd (PF10613 ,Canal iónico ligado L-glutamato y sitio de unión a glicina, 557-661); ANF_receptor (PF01094, Región de unión del ligando de la familia de receptores, 124-471).

InterPro proporciona análisis funcional de proteínas clasificándolas en familias y prediciendo dominios y sitios importantes. Para clasificar las proteínas de esta manera, InterPro utiliza modelos predictivos (firmas), proporcionados por diversas bases de datos. Para nuestra proteína de interés se encontró la siguiente información: Receptor NMDA 3A, subtipo de receptor NMDA de canales iónicos activados por glutamato con conductancia monocanal reducida, baja permeabilidad al calcio y baja sensibilidad al magnesio dependiente del voltaje. Mediada por glicina. Durante el desarrollo de los circuitos neuronales, juega un papel en el período de refinamiento sináptico, restringiendo la maduración y el crecimiento de la columna. Al competir con la interacción GIT1 con ARHGEF7 / beta-PIX, puede reducir la activación local regulada por GIT1 / ARHGEF7 de RAC1, lo que afecta la señalización y limita la maduración y el crecimiento de las sinapsis inactivas. También puede desempeñar un papel en el mecanismo de señalización mediado por PPP2CB-NMDAR. (Protein family membership: receptor ionotrópico de glutamato, metazoa, IPR001508).

En la base de datos CDD, las secuencias de proteínas de estructuras tridimensionales se incluyen en modelos de dominio siempre que sea posible, uno de los objetivos es hacer que las alineaciones de secuencias múltiples estén de acuerdo con lo que podemos inferir de la estructura tridimensional y la superposición de estructuras tridimensionales, para comprender las relaciones secuencia / estructura / función (Figura 14).

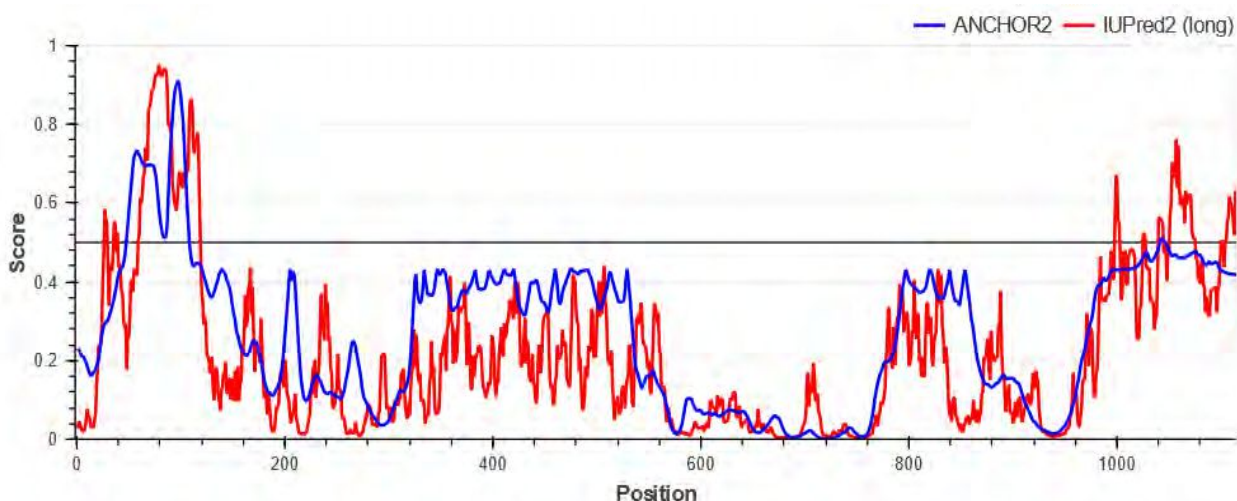


Figura 13. Regiones desordenadas predichas por IUPred2A. Las regiones desordenadas son aquellos picos que superan el valor de score delimitado con una línea negra. Se observan 2 regiones de mayor desorden: segmento que va de la posición 40 a la 160 aproximadamente, y un segundo segmento ligeramente desordenado que va de la posición 960 (aprox) a la 1115. Además se puede observar que aquellas regiones que se clasificaron como transmembrana (600-800 y 800-900) son las regiones con menos desorden, más conservadas y predecibles. También podemos ver que las regiones de unión desordenadas identificadas por el programa tienen un pico en el primer segmento de la secuencia (1-200), que a su vez coincide con la región que contiene el segmento más desordenado de la proteína. Se utilizaron los parámetros por default: IUPred2 long disorder (default).

El Conserved Domain Architecture Retrieval Tool (CDART) encuentra similitudes proteicas a través de distancias evolutivas significativas utilizando perfiles de dominio sensibles en lugar de similitud de secuencia directa. Dada una secuencia proteica de consulta, CDART muestra los dominios conservados que la componen, identificados por RPS-BLAST, y luego enumera las proteínas con una arquitectura de dominio conservado similar. Realizando una búsqueda con parámetros por default con la secuencia del receptor ionotrópico NMDA 3A se encontraron un total de 3161 arquitecturas, donde los primeros 3 hits contenían 3 dominios (Figura S12) :

-Periplasmic binding protein type 1 (intervalo 40-496). Compuesto por la familia de reguladores transcripcionales de los dominios de unión a Lacl, proteínas periplasmáticas del transporte ABC, la familia de receptores GPCRs, y familia de receptores NPRs.

-Lig Chan superfamily (intervalo 676-942). Este dominio incluye las 4 regiones transmembranales de los receptores ionotrópicos NMDA.

-Periplasmic binding protein type 2 (intervalo 512-908). Representa los dominios de unión y ligamiento encontrados en proteínas ligadoras de solutos, que sirven como receptores iniciales en el transporte, transducción de señales, y channel gating.

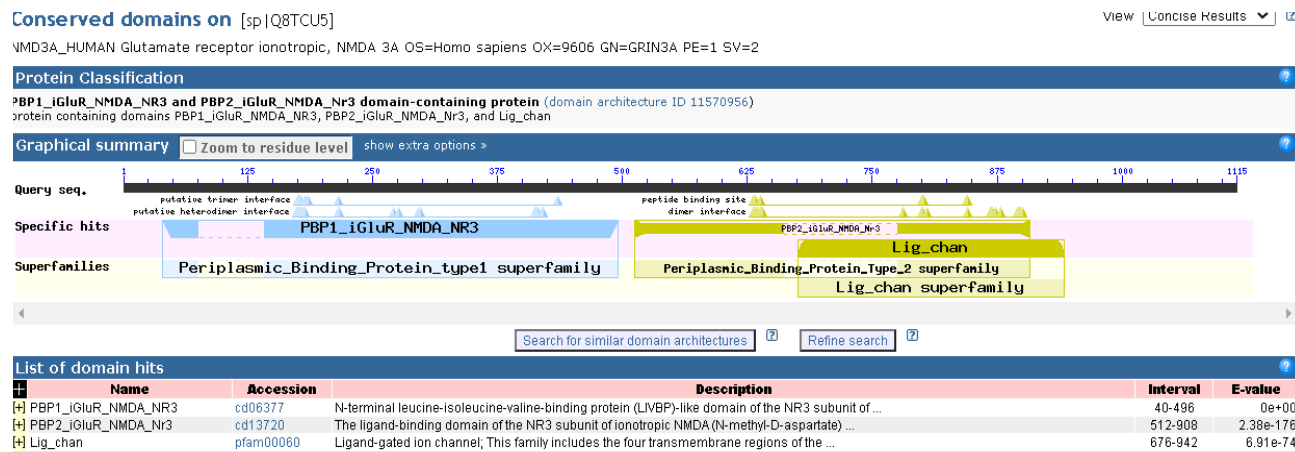


Figura 14. Clasificación dada por CCD para la proteína Q8TCU5. La proteína pertenece a una superfamilia de canales iónicos de ligando controlados que presenta 4 regiones transmembrana de los receptores ionotrópicos de glutamato y los receptores NMDA. También se destaca la superfamilia de pliegues de unión periplásmica tipo 2; este modelo evolutivo y jerarquía representan los dominios de unión de ligandos que se encuentran en las proteínas de unión de solutos que sirven como receptores iniciales en el transporte, la transducción de señales y la activación de canales. El origen del módulo PBP se puede rastrear a través de los filos distantes, incluidos eucariotas, arqueobacterias y procariontes. Además de las proteínas de transporte, la familia incluye receptores de glutamato ionotrópicos y proteínas sensoras no ortodoxas implicadas en la transducción de señales. El dominio de unión al sustrato de los reguladores de la transcripción LysR y los sistemas de transporte de tipo oligopéptido también contienen el pliegue de unión periplásmico de tipo 2 y, por tanto, son significativamente homólogos al de PBP2. Parámetros: Expect Value: 0,01, Maximum number of hits: 500.

Se realizó la búsqueda de motivos secuenciales mediante ScanProsite (De Castro, et. al 2006), un servidor que permite escanear proteínas en busca de coincidencias con la colección de motivos PROSITE, así como con patrones definidos por el usuario. Como resultado del escaneo rápido con parámetros por default (excluyendo motivos con alta probabilidad de ocurrencia) no se encontraron motivos. Sin embargo, al incluir en la búsqueda motivos con alta probabilidad de ocurrencia se obtuvieron los resultados mostrados en la figura 15.

Por último, para obtener la clasificación estructural de Q8TCU5 se realizó una búsqueda secuencial en CATH (Greene, et. al 2007). La base estructural CATH clasifica dominios proteicos basándose en comparación estructural. Los dominios que comparten los primeros 4 números de CATH son homólogos. Esta base de datos está organizada de forma jerárquica, con 9 clasificaciones: clase (proporción de residuos que adoptan la conformación α -hélice o β -plegada), arquitectura (estructura secundaria en el espacio), topología (conectividad y arreglo de la estructura secundaria), superfamilia de homólogos (dominios homólogos), familia secuencial, familia de ortólogos, dominios similares, dominios idénticos, y dominios únicos. Los primeros 4 niveles corresponden a una clasificación del tipo secuencial y estructural, las restantes se basan únicamente en una clasificación secuencial. Se encontraron 84 coincidencias en la búsqueda secuencial de dominios (principalmente regiones de unión a ligandos, desactivación de los sitios activos, etc), y unos 50

resultados en la búsqueda secuencial de familias funcionales (mayormente familias del receptor NMDA, AMPA, y kainato) (Figura 16).

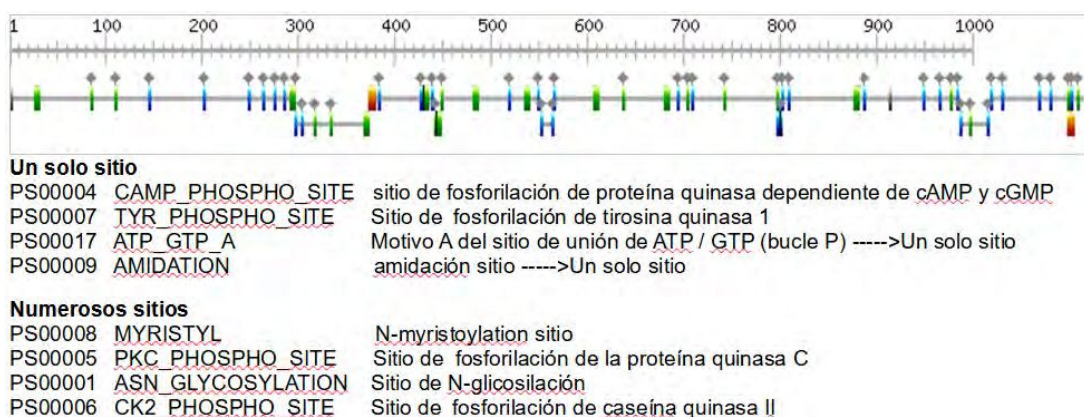


Figura 15. Motivos con alta probabilidad de ocurrencia detectados. La figura representa aciertos por patrones con una alta probabilidad de ocurrencia o por patrones definidos por el usuario: [61 aciertos (por patrones distintos) en 1 secuencia]. Los motivos con una alta probabilidad de aparición son en la mayoría de los casos patrones que se encuentran en muchas secuencias de proteínas. Algunos de ellos describen, modificaciones postraduccionales que se encuentran comúnmente, y algunas otras regiones con sesgo de composición.

	Level	CATH Code	Description
	3		Alpha Beta
	3.40		3-Layer(aba) Sandwich
	3.40.190		D-Maltodextrin-Binding Protein; domain 2
	3.40.190.10		Periplasmic binding protein-like II

αβ 3-Layer Sandwich(aba) (1ntr)

Figura 16. Clasificación dada por CATH. En el recuadro se muestra las categorías con las que se clasificó a la proteína de acuerdo a la estructura jerárquica de CATH: clase=alfa-beta, arquitectura=3-layer(aba) Sandwich, topología= proteína de unión a D-Maltodextrina; dominio 2.

Asignación de plegamiento

El primer paso del proceso de modelado comparativo está definido como la predicción del plegamiento estructural de la proteína objetivo a partir de su secuencia de aminoácidos mediante la detección de proteínas homólogas de estructura tridimensional conocida. La idea se basa en la hipótesis de Anfinsen de que la estructura de las proteínas en un determinado entorno es a su vez determinada por la secuencia de proteínas. Una vez que se predice el plegamiento de la proteína objetivo, se debe elegir el mejor candidato a utilizar como template, cuya óptima elección es fundamental para asegurar la calidad del modelo a obtener.

A continuación se describen los métodos y bases de datos utilizadas para obtener los homólogos estructurales más cercanos de la proteína GluN3A.

En una primera instancia, se realizó la búsqueda de homólogos en la base de datos PDB (Burley, et. al 2021) (Protein Data Bank). PDB organiza las macromoléculas biológicas por su estructura jerárquica (estructura primaria, secundaria, terciaria, y cuaternaria) para simplificar la búsqueda, proporcionando sus respectivas estructuras obtenidas por técnicas de difracción de rayos X, RMN, microscopía crioelectrónica y modelado teórico. Como resultado no se encontraron entradas específicas del receptor ionotrópico de glutamato NMDA subunidad 3A de humanos, pero sí se encontraron entradas de estructuras homólogas con hasta un 50% ID. De los resultados de la búsqueda por similitud secuencial en la base PDB se obtuvo la siguiente lista

de proteínas: 2RC9_1 (69%ID), 4KCD_1 (69%ID), 2RC7_1 (69%ID), 2RC8_1 (69%ID), 2RCA_1 (57%ID), 2RCB_1 (57%ID). Los resultados se resumen en la tabla 2.

Con el objetivo de buscar diferentes conformeros estructurales se utilizó la base de datos CoDNaS (Monzon, et. al, 2016). La base de datos CoDNaS es una colección de estructuras cristalográficas redundantes para una proteína dada ampliamente vinculada con información estructural, biológica y fisicoquímica. Varias proteínas depositadas en la base de datos PDB se han cristalizado en diferentes condiciones (por ejemplo con y sin la presencia de un ligando dado, en diferentes estados oligoméricos, con o sin presencia de modificaciones postraduccionales, etc.), que según la teoría de la selección conformacional, son factores que se podrían usar para estudiar los cambios conformacionales y correlacionarlos con la información biológica. Al realizar la búsqueda con la secuencia de la proteína GluN3A no se encontraron resultados. Esto se debe a que esta proteína no está anotada en la base PDB, debido a que su estructura aún no fue dilucidada. En cambio, al buscar el Código de la proteína 2RC9 (homólogo estructural más cercano), la base de datos arrojó un único resultado correspondiente a la proteína 2CR7, que resulta ser otra conformación de la proteína 2RC9, ya que comparten el mismo Código de Uniprot (Bateman, 2021) [Q9R1M7](#), cuyo nombre de entrada es NMD3A_RAT.

Adicionalmente se realizó la búsqueda de homólogos con estructura conocida mediante Psi-blast, HHPred (Zimmermann, et. al, 2018), LOMETS (Wu S, y Zhang Y, 2007) y Phyre2 (Kelley, et. al, 2015). En el primer caso los primeros 8 hits fueron identificados como posibles template por sus elevado %ID y query coverage, bajo valor de e-value, y por poseer una estructura definida en la base PDB (Tabla 4).

Tabla 4. Búsqueda de homólogos de estructura conocida.

Accession number	Código PDB	Código Uniprot
NMDZ1_RAT	4KCC	P35439
GRIA2_RAT	4YU0	P19491
NMD3A_RAT	2RC7	Q9R1M7
NMDE2_RAT	5FXG_B	Q00960
NMDZ1_RAT	6WHR_B	P35439
NMDE2_RAT	6CNA_B	Q00960
NMDZ1_RAT	6WI1_B	P35439
NMDZ1_XENLA	5UOW_D	A0A1L8F5J9

HHpred es un método para la búsqueda de bases de datos de secuencias y la predicción de estructuras que es mucho más sensible que BLAST o PSI-BLAST para encontrar homólogos remotos. La secuencia objetivo o MSA se utiliza para construir un HMM (Hidden Markov Model), que está alineado con todos los HMM que representan proteínas anotadas o dominios con estructura conocida en bases de datos de alineación como Pfam y SMART. Los HMM pueden incluir información SS (experimentalmente determinada o predicha). Como resultado se obtuvieron 377 hits, con una Query MSA diversity (Neff) de 6.28031. Los 10 primeros hits de la búsqueda se muestran en la tabla 5.

Adicionalmente, se realizó una búsqueda con HHPred utilizando el MSA de los 100 primeros hits obtenidos por PSI-BLAST con, los siguientes parámetros por defecto: MSA generation method=HHblits=>UniRef30, MSA generation iterations=3, E-value cutoff for MSA generation=1e-3, Min seq identity of MSA hits with query (%)=20, Min coverage of MSA hits (%)=40, Secondary structure scoring=during_alignment, Alignment Mode: Realign with MAC =global:realign, MAC realignment threshold=0.3, Max target hits=1000, Min probability in hitlist (%)=20. (Tabla 6).

Tabla 5. Resultados de la búsqueda de homólogos remotos, de estructura y/o dominios conocidos mediante HHpred.

Uniprot accession number	Código Uniprot	Código PDB
GRIK5_RAT	Q63273	7KS0_D
GRIA1_MOUSE	P23818	7LDD_C
NMDZ1_RAT	P35439	6W11_A
GRIK5_RAT	Q63273	7KS0_C
GRIK3_RAT	P42264	6JFY_C
NMDZ1_XENLA	A0A1L8F5J9	4TLL_C
P96404_MYCTU	P96404	6LDZ_A
NMDZ1_XENLA	A0A1L8F5J9	4TLL_D
GRID2_RAT	Q63226	6LU9_C
GRIA2_RAT	P19491	6PEQ_A

Tabla 6. MSA de los primeros 100 hits obtenidos por PSI-BLAST.

Uniprot accession number	Código Uniprot	Código PDB
GRIK5_RAT	Q63273	7KS0
NMDZ1_RAT	P35439	6W11_A
NMDZ1_XENLA	A0A1L8F5J9	4TLL_D
GRIA1_MOUSE	P23818	7LDD_C
NMDZ1_XENLA	A0A1L8F5J9	4TLL_C
GRIK3_RAT	P42264	6JFY_C
NMDZ1_XENLA	A0A1L8F5J9	5TQ0_B y 5UN1_F
P96404_MYCTU	P96404	6DLZ_A
GRID2_RAT	Q63226	6LU9_C

LOMETS (Local Meta-Threading Server, versión 3) es un predictor de la estructura de proteínas basada en templates y la anotación de funciones basada en la estructura, que integra múltiples métodos de subprocesamiento basados en el deep-learning (CEthreder, DisCover, EigenThreader, Hybrid-CEthreder, MapAlign) y programas basados en perfiles de última generación (FFAS3D (Xu D, 2013), HHpred, HHsearch, MRFsearch, MUSTER, SparksX). Con parámetros por default se identificaron 10 posibles templates, de los cuales se seleccionaron 10 de ellos para realizar el modelado final de la proteína con MODELLER (Webb B, & Sali A, 2016) y se utilizó L-BFGS para objetivos no homólogos mediante las restricciones de distancia predichas por DeepPotential y calculadas a partir de los templates mejores (top templates).

Phyre2 es un conjunto de herramientas disponibles en la web para predecir y analizar la estructura, función y mutaciones de las proteínas. Utilizando métodos avanzados de detección de homología remota. PHYRE2 construye modelos 3D, predice sitios de unión de ligandos y analiza el efecto de variantes de aminoácidos (por ejemplo, SNP no sinónimos (nsSNP)) para la secuencia de proteínas de un usuario. Proporcionando una secuencia de proteínas se puede obtener: la interpretación de la estructura secundaria y terciaria de sus modelos, la composición de su dominio y la calidad del modelo. La búsqueda de template con este servidor arrojó 6 posibles templates: c6irfD, c5uowB, c6mmiD, c4pe5B, c5kbuA, c4pe5A. La predicción de PHYRE2, con parámetros por defecto, fue realizada con 6 templates con base en métodos heurísticos para maximizar la confianza, el %ID, y el query coverage. El modelo predicho se presenta en el apéndice (Figura S13). Si bien el modelado fue intensivo, hay dos fragmentos sin modelar ya que los templates elegidos no cubren esos fragmentos. Las regiones sin modelar son (1-175) aproximadamente, y la (955-1115), ya que se observan desprovistos de una estructura secundaria determinada por falta de información en el template para su modelado. Hubo un total de 293 posiciones que fueron modeladas ab initio (tener en cuenta que el

modelado ab initio tiene bajo nivel de confianza), por falta de un template adecuado que cubriera estos segmentos. Los candidatos a template se encuentran resumidos en la tabla 7.

Tabla 2. Búsqueda por similitud secuencial en la base PDB del receptor ionotrópico del glutamato subunidad 3A.

ID de proteína	Descripción	Ligando
4KCD	Crystal Structure of the NMDA Receptor GluN3A Ligand Binding Domain Apo State. <i>Rattus norvegicus</i>	Glicerol
2RC9	Crystal structure of the NR3A ligand binding core complex with ACPC at 1.96 Angstrom resolution. <i>Rattus norvegicus</i>	ácido 1-aminociclopropano carboxílico
2RC7	Crystal structure of the NR3A ligand binding core complex with glycine at 1.58 Angstrom resolution. <i>Rattus norvegicus</i>	Ion bromuro
2RCB	Crystal structure of the NR3B ligand binding core complex with D-serine at 1.62 Angstrom resolution. <i>Rattus norvegicus</i>	D-serina
2RC8	Crystal structure of the NR3A ligand binding core complex with D-serine at 1.45 Angstrom resolution. <i>Rattus norvegicus</i>	Ion cloruro
2RCA	Crystal structure of the NR3B ligand binding core complex with glycine at 1.58 Angstrom resolution. <i>Rattus norvegicus</i>	Glicina

Tabla 3. Resultados de la búsqueda del Código de Uniprot 2RC9.

ID_POOL_ CoDNAs	UniProt	#CONF	RMS D min	RMSD max	RMSD avg	Protein Name
2RC7_A	Q9R1M 7	8	0.2	1.35	0.7682	Glutamate NMDA receptor subunit 3A

Nota: código correspondiente al homólogo estructural más cercano del que se tiene información, en relación a la proteína Q8TCU5.



Figura 17. Modelo final de LOMETS. Construido con Modeller con los 10 templates (detectados por TM-align), estos análogos estructurales no logran una buena cobertura de la proteína ya que apenas logran cubrir un 58% de ella. Los alineamientos estructurales son de baja resolución en general, solo pudiéndose obtener hasta un 4,7 Å de RMSD para el mejor alineamiento estructural (con la proteína 3KG2), además de no poder lograr altos porcentajes de identidad entre los candidatos a template (%ID<20) y la proteína Q8TCU5, lo que posiblemente sea porque hay dominios de la proteína no identificados en las bases de datos, y por lo tanto no cubiertos.

Analizando los parámetros de la tabla 7 se seleccionó como template a la proteína 7ks0 con sus respectivas cadenas (A, B, y C), por su gran porcentaje de cobertura e identidad con la proteína Q8TCU5, además de poseer un score de alineamiento razonable (>800), siendo identificado como el homólogo estructural más cercano mediante la base RCSB PDB.

Modelado por homología utilizando Modeller

Teniendo en cuenta que las proteínas relacionadas evolutivamente tienden a tener estructuras similares, muchas estructuras proteicas pueden modelarse basándose en estructuras ya conocidas de proteínas relacionadas, a esto se le llamó “modelado por homología”. Los modelos de la proteína objetivo se obtienen basándose en las coordenadas atómicas de estructuras conocidas, bajo el supuesto de que la estructura de la proteína a modelar es similar a la estructura de la proteína utilizada como template (estructura que servirá de molde para el modelado del objetivo), donde además la proteína objetivo y el template están relacionados evolutivamente (siendo proteínas homólogas), y en el alineamiento entre el template elegido y el objetivo. Uno de los programas más utilizados es Modeller, que modela estructuras tridimensionales de proteínas y sus ensamblajes mediante satisfacción de las restricciones en la estructura espacial de la(s) secuencia(s) de aminoácidos y ligandos a modelar, resultando en una estructura 3D que satisface estas restricciones lo mejor posible. El modelo 3D se optimiza con la función de densidad de probabilidad molecular (molpdf). Para el modelado molecular de la proteína NMD3A_HUMAN (Q8TCU5) los template elegidos fueron: 7KS0_C (del aminoácido 33-845) y 7KS0_C (del aminoácido 21-829). Como resultado se obtuvieron 10 modelos de los cuales, en base al mínimo valor de tanto el DOPE score, como el valor de Molpdf, se eligió el modelo UKNP.B99990006.

Tabla 8. Parámetros utilizados para seleccionar el modelo.

Nombre del modelo	Mol pdf	DOPE score
UKNP.B99990006	7197.50928	-91363.32813

El modelo UKNP.B99990006 fue posteriormente optimizado para obtener un mejor RMSD, removiendo los átomos que no pudieron ser modelados, el primer fragmento removido va del 1 al 44, y el otro fragmento removido va de la posición 957 a la 1016 en el modelo final. Se utilizó Pymol (visor y renderizador molecular) para comparar estructuralmente los templates utilizados contra el modelo obtenido de la proteína objetivo, y se optimizó el alineamiento estructural al remover los átomos no modelados del modelo final, disminuyendo el RMSD del alineamiento. Se obtuvo un RMSD=4.103 para 7KS0_D con el modelo, y un RMSD=21,494 para 7KS0_C y el modelo (Figura S14 (a) y (b)).

Para evaluar la calidad del modelo obtenido se utilizó ProSA (Wiederstein & Sippl, 2007), ya que calcula un puntaje de calidad general para una estructura de entrada específica. Si esta puntuación está fuera de un rango característico de las proteínas nativas, es probable que la estructura contenga errores. Para el análisis del modelo UKNP.B99990006 obtenido con Modeller se obtuvieron los que se muestran en la figura 18 (a), (b), y (c).

Para el análisis funcional se utilizó nuevamente Pymol para observar la superficie de la proteína modelada y sus interacciones (Figura S15).

Tabla 7. Selección de template.

Candidato a template	Codigo PDB	Resolución (Å)	Identidad	Query coverage	Score alignent (T-coffe)
GRIK5_RAT	7ks0_C	5,3	22%	33%	804
GRIK5_RAT	7ks0_A	5,3	22%	33%	804
GRIK5_RAT	7ks0_D	5,3	40%	40%	845
GRIK5_RAT	7ks0_B	5,3	40%	40%	845
NMDZ1_RAT	4pe5_A	3,96	24%	54%	878
GRIA2_RAT	3kg2_A	3,60	21%	33%	786
GRIA1_MOUSE	7ldd_A	3,40	20%	72%	850

Nota: Para la elección de los templates adecuados se utilizó como criterio los parámetros que se muestran en la tabla, priorizando aquellas entradas pdb que cuentan con una resolución más cercana a 2 Å, contando con un buen coverage de la proteína objetivo, y un alto valor para su porcentaje de identidad. Se han suprimido los candidatos que contaban con un gran número de gaps (como por ejemplo la proteína Suow).

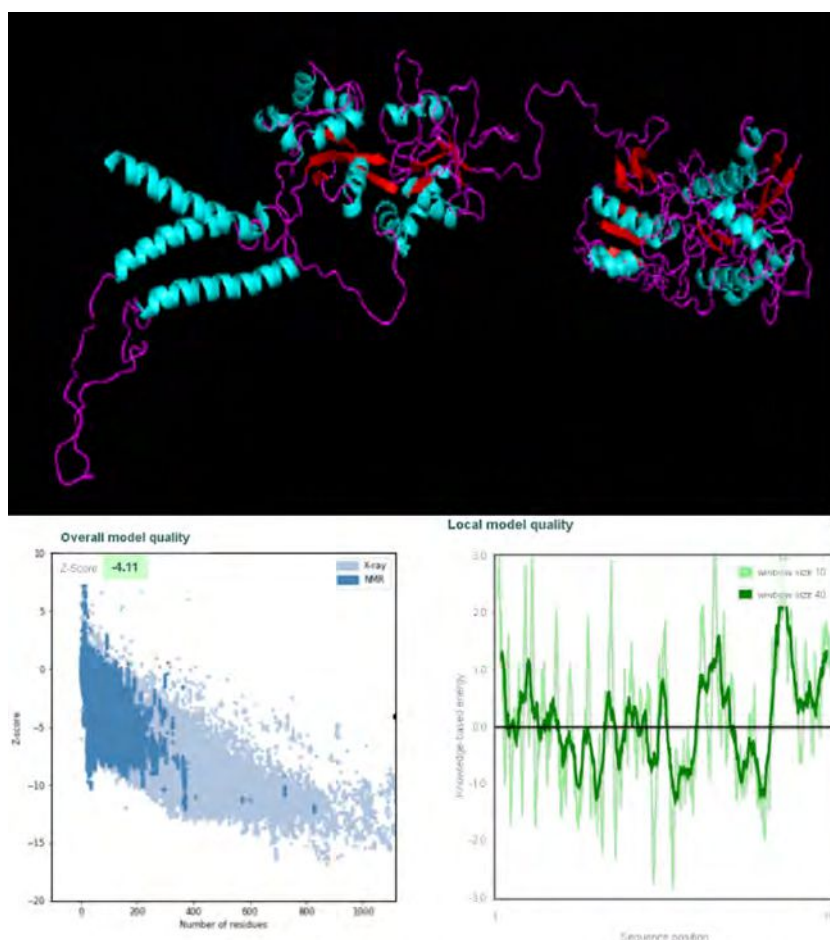


Figura 18 (a). Modelo refinado de Q8TCU5. En la figura se observa la estructura del modelo UKNP.B99990006 refinado con su correspondiente estructura secundaria: en color celeste se indican las hélices alfa, en color rojo se indican las hojas beta plegadas, y en fucsia los loops.(b). Calidad general del modelo. La figura izquierda representa la calidad general del modelo en función de los valores del Z-score obtenidos para estructuras determinadas experimentalmente (rayos X, NMR), como el valor de Z-score obtenido para el modelo UKNP.B99990006 (punto negro con un Z-score= -4.11) esta muy alejado del rango de valores para las estructuras experimentales que comparten la misma longitud para su secuencia, se puede decir que el modelo obtenido contiene grandes fallas. (c). Calidad del modelo local. La figura de la derecha muestra la calidad del modelo local, donde la línea color verde oscura simboliza la energía promedio sobre cada fragmento de 40 residuos, mientras que la línea de color verde claro representa la energía promedio sobre cada fragmento de 10 residuos. En general, los valores positivos corresponden a partes problemáticas o erróneas de la estructura de entrada. De esta manera se puede observar cuáles regiones del modelo obtenido presentan mayores problemas de modelado, en este caso se puede observar que solo una pequeña fracción de la proteína fue correctamente modelada (solo dos dominios) ya que prevalecen los picos de valores positivos.

Estimación filogenética

Como primera etapa en el estudio filogenético se obtuvo un alineamiento múltiple con T-Coffee (Notredame, et. al 2000) (Version_11.00, Cedric Notredame) utilizando secuencias de las siguientes

especies: *Homo sapiens*, *Pan troglodytes*, *Gorilla gorilla gorilla*, *Theropithecus gelada*, *Mandrillus leucophaeus*, *Papio anubis*, *Chlorocebus sabaeus*, *Cercocebus atys*, *Macaca Nernestrina*, *Macaca fascicularis*, *Hylobates moloch*, *Propithecus coquereli*, *Rhinopithecus roxellana*, *Rhinopithecus bieti*, *Colobus angolensis palliatus*, *Pongo abelii*, *Ptilocolobus tephrosceles*, *Nomascus leucogenys*, *Trachypithecus francoisi*, *Aotus nancymae*, *Cebus imitator*, *Saimiri boliviensis boliviensis*, *Sapajus apella*, *Callithrix jacchus*, *Rhinolophus ferrumequinum*, *Microcebus murinus*, *Otolemur garnettii*, *Carlito syrichta*. (N° de secuencias alineadas=35), (Figura S16).

Analizando el receptor ionotrópico del glutamato NMDA subunidad 3A, perteneciente a especies a la orden de los primates, se encontró una gran conservación de la proteína en general, siendo la proteína mayormente conservada, logrando un 99% ID promedio, y con un query coverage del 100%. Únicamente en el bloque n°12 que comprende el intervalo (1000-1080), hubo mayores diferencias con un total de 9 posiciones similares pero no idénticas, que es una de las regiones que presenta mayor desorden. En cuanto el resto de la proteína las variaciones por bloque eran mínimas (solo posiciones variantes en Q8TCU5) , lo que indica una gran conservación de la proteína receptora del glutamato (subunidad 3A) en primates.

A continuación se utilizó Modeltest es un subprograma del software HYPHY (Pond, 2005) para seleccionar el modelo de proteínas que mejor representa las secuencias bajo estudio. Por medio de un análisis estándar se generó una reconstrucción filogenética con las 35 secuencias homólogas de primates obtenidas con Blast, por el método de distancia (calculando una distancia genética entre un par de especies) de Neighbour Joining. Con este árbol se realizó una comparación entre modelos, obteniéndose como mejor modelo HIV between+F. Finalmente, se utilizó el software PHYML (Guindon, & Gascuel, 2003) para la inferencia de la filogenia basado en el principio estadístico de Maximum likelihood (Parámetros utilizados: Model of amino-acids substitution=HIVw, Amino-acids frequencies=empirical, Number of substitution rate=4, Gamma distributed rate across sites=yes, Tree topology search options= Best of NNI and SPR, Non parametric bootstrap analysis=yes, -Number of replicates=100, el resto de los parámetros se utilizaron por defecto) (Figura 19).

Predicción de función

Se realizó una búsqueda bibliográfica acerca del receptor ionotrópico de glutamato NMDA subunidad 3A. En la tabla 9 se resume parte de la información obtenida.

Se utilizó el servidor ConSurf (Ashkenazy, et. al 2016) para estimar la conservación evolutiva de aminoácidos basada en las relaciones filogenéticas entre secuencias homólogas. El grado en el que una posición se conserva evolutivamente (tasa de evolución estimada por método bayesiano o método de máxima verosimilitud) depende en gran medida de su importancia estructural y funcional, por lo que el análisis de conservación de posiciones entre miembros de la misma familia a menudo puede revelar la importancia de cada posición para la estructura o función de la proteína. Se obtuvieron múltiples output de MSA que mostraban la conservación o variabilidad de aminoácidos en las 35 secuencias de primates analizadas, así como también se obtuvo un árbol filogenético, obtenido por ML (Figura 20).

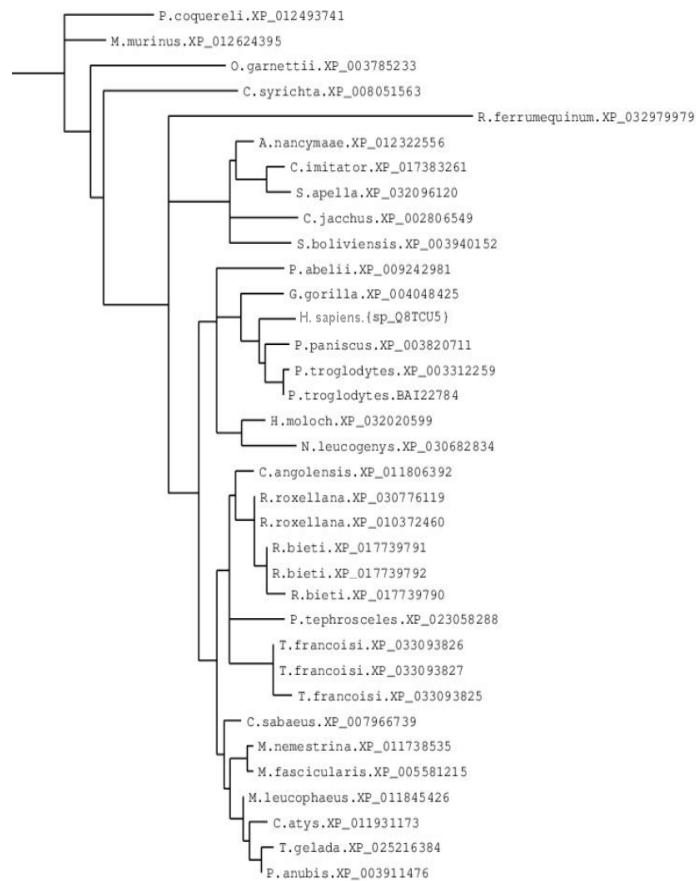


Figura 19. Árbol filogenético obtenido con PhyML. En la figura se muestra el árbol roteado obtenido (con los 35 homólogos cercanos obtenidos anteriormente en Blast) en PhyML y posteriormente modificado con SNAD para una mejor visualización, se analizó la divergencia evolutiva del receptor ionotrópico del glutamato NMDA 3A. Se observa que la proteína NMDA GRIN3A de *H. sapiens*, tiene una distancia evolutiva menor con la proteína expresada en *P. paniscus* (chimpancé gracil), y con la proteína expresada en *P. troglodytes* (chimpancé común), lo cual es esperable ya que son las especies evolutivamente más cercanas a los *H. sapiens* a nivel genético. La especie con mayor divergencia secuencial respecto a la proteína encontrada en *H. sapiens* es *Papio anubis*.

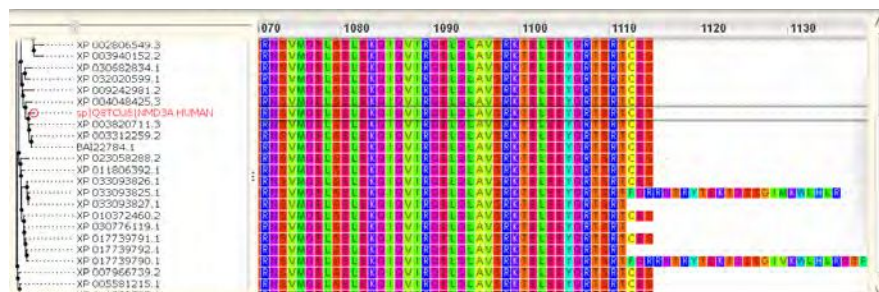
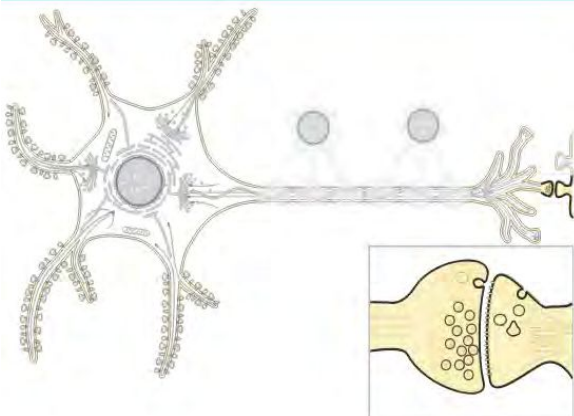


Figura 20. Árbol filogenético construido por ConSurf visualizado con Wasabi (Veidenberg, et al, 2016). Se observa en la figura la variabilidad secuencial entre las secuencias de primates, donde existen dos segmentos variables no compartidos entre primates, y a su vez estos segmentos terminales tienen mayor longitud que en el resto de los primates. En el margen izquierdo se muestra el árbol construido por máxima similitud.

Tabla 9. Información general acerca del receptor ionotrópico NMDA 3A.

Nombres:	Glutamate receptor ionotropic, NMDA 3A/ GluN3a/NMDAR3A/NR3A
Función molecular	Canal de iones , canal iónico de apertura por ligando , Receptor
Proceso biológico	Transporte de iones , Transporte
Ligando	Calcio , magnesio
Nombres de genes	Nombre: GRIN3A
Organismo	Homo sapiens (humano)
Identificador taxonómico	9606 [NCBI]
Linaje taxonómico	Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Mammalia › Eutheria › Euarchontoglires › Primates › Haplorrhini › Catarrhini › Hominidae › Homo
Proteomas	UP000005640 Componente i : Cromosoma 9
Función molecular	Actividad del canal iónico controlado por ligando. Fuente: GO_Central Actividad del receptor de glutamato NMDA Fuente: UniProtKB Unión a proteína fosfatasa 2A. Fuente: UniProtKB Actividad del receptor de señalización. Fuente: GO_Central Actividad del canal iónico dependiente del transmisor involucrado en la regulación del potencial de membrana postsináptica. Fuente: Ensembl Transporte de iones de calcio Fuente: UniProtKB Desarrollo de dendrita Fuente: Ensembl Vía de señalización del receptor de glutamato ionotrópico Fuente: GO_Central Regulación negativa del desarrollo de la columna dendrítica Fuente: UniProtKB Inhibición prepulso Fuente: Ensembl Regulación de la exocitosis de vesículas sinápticas Fuente: Ensembl Respuesta al etanol
Ubicación (Uniprot)	subcelular  <div style="display: flex; flex-direction: column; align-items: flex-end; margin-top: 10px;"> <div style="display: flex; gap: 5px;"> Membrana celular ⓘ Por similitud </div> <div style="display: flex; gap: 5px;"> membrana de múltiples pasadas ⓘ Por similitud </div> <div style="display: flex; gap: 5px;"> membrana celular postsináptica ⓘ Por similitud </div> <div style="display: flex; gap: 5px;"> densidad postsináptica ⓘ Por similitud </div> <div style="font-size: small; margin-top: 5px;"> <p><i>Nota:</i> Enriquecido en membrana plasmática densidades postsinápticas. Requiere la pres apuntar a la membrana plasmática (por sim) Por similitud</p> </div> </div>

Para tratar de establecer una función para el dominio C-terminal se utilizó PSIPRED. Al realizar una búsqueda mediante este servidor para el fragmento C-terminal de la proteína Q8TCU5, es decir, de la posición 750 a la 1115, para el cual no se pudieron obtener predicciones fiables, se reconocieron dos posibles dominios en la región c-terminal (Figura 5, 6, 7, 31, y 32). De esta forma se detectó un segmento terminal extracelular.

Para la predicción de la función de la proteína objetivo se utilizó un método de aprendizaje automático, el método Support Vector Machine (Cai, et. al, 2001) para predecir la clase funcional de las proteínas y péptidos a partir de sus propiedades secuenciales, independizándose de la similitud secuencial. Las predicciones concuerdan ampliamente con las derivadas de InterPro, pudiendo resumir los resultados en la tabla 10.

CONCLUSIONES Y DISCUSIÓN

Se puede decir, dado los sucesivos análisis, que el receptor NMDA GRIN3A está altamente conservado en primates, posee al menos 4 hélices transmembrana, además de tener una región coiled-coil hacia el final de su secuencia (1000-1115) que es un segmento sin dominios detectables por su alta variabilidad secuencial, además de ser una de las regiones de mayor desorden junto con el segmento que va del 1 al 200 (posible

péptido señal). El resto de la proteína tiene una estructura rígida, por ende está bien conservada por lo que debe ser una proteína con una función biológica bien definida.

Tabla 10 . Resumen de los GO terms más relevantes.

Proceso biológico	Función molecular	Componente celular
Transporte de iones (GO:0006811)	actividad del receptor de glutamato ionotrópico (GO: 0004970)	membrana (GO: 0016020)
proceso biosintético de macromoléculas celulares (GO:0034645)	actividad del canal iónico controlado por ligando (GO: 0015276)	componente intrínseco de la membrana (GO:0031224)
Vía de señalización del receptor de superficie celular (GO:0007166)	actividad del canal iónico (GO: 0005216)	componente integral de la membrana (GO:0016021)
transporte de iones transmembrana (GO:0034220)	actividad del receptor de señalización (GO: 0038023)	periferia celular (GO:0071944)

Nota: Resumen de los GO terms asociados más relevantes recopilados de Interpro, QuickGO, Uniprot, y el software AmiGO de GeneOntology (GO).

Para el alineamiento múltiple de secuencias con la herramienta T-Coffee se seleccionaron 35 secuencias c/u perteneciente a las especies del orden de los primates, donde las especies con menor distancia evolutiva son: Homo sapiens, Pan troglodytes, Gorilla gorilla gorilla. La proteína está bien conservada, por lo que seguramente cumplirá la misma función biológica en estos 3 organismos (por transferencia). Se encontró en el MSA que los segmentos de las secuencias alineadas que muestran mayor número de gaps (mayor número de mutaciones, y por ende menor conservación) son los que corresponden a los segmentos de mayor desorden en NMD3A _ HUMAN (donde se predice una hélice transmembrana). Se puede observar que hay tres especies de primates con una región C-terminal diferente: *Carlito syrichta* (especie más alejada evolutivamente del resto de los primates), *Trachypithecus francois*, y *Rhinopithecus bieti* que poseen una región C-terminal de mayor longitud que la esperada. Todas las secuencias de homólogos analizadas pertenecen al mismo cluster por guardar una alta similitud secuencial, estructural, y funcional.

No se han encontrado unidades de repetición por lo que no hay evidencias para decir que esta sea una proteína repetitiva. Al buscar dominios conservados se encontraron 4 dominios: la familia Lig_chan que es un receptor ionotrópico de glutamato en el segmento (674-942), la familia Lig_chan-Glu_bd que es una región con canal iónico ligado L-glutamato y sitio de unión a glicina en el segmento (557-661), la familia ANF_receptor que es una región de unión del ligando de la familia de receptores en el segmento (124-471), y el dominio de proteínas de unión periplásmica tipo 1 y 2 (60-690) aproximadamente. Solo se encontraron motivos secuenciales que se encuentran en una gran mayoría de proteínas por ser muy comunes. Se observó una gran cantidad de regiones de baja complejidad en el fragmento N-inicial y el C-terminal de la secuencia de la proteína Q8TCU5, lo que implica que estas dos regiones son fuente de alta variabilidad genética, propensas a aumentar la frecuencia de desarrollo de enfermedades neurodegenerativas, y podrían estar implicados en los procesos de adaptación (evolución del cerebro social humano).

En cuanto a la función molecular es un receptor de NMDA de canales iónicos activados por glutamato, de baja permeabilidad al calcio, por lo cual se predice que es un componente intrínseco de la membrana. Su función biológica está vinculada al refinamiento sináptico, restringiendo la maduración y el crecimiento de la columna, así como se relaciona a la poda de las sinapsis inactivas.

BIBLIOGRAFÍA

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-3402.
- Andres Veidenberg, Alan Medlar, Ari Löytynoja, Wasabi: An Integrated Platform for Evolutionary Sequence Analysis and Data Visualization. *Molecular Biology and Evolution*, Volume 33, Issue 4, April 2016, Pages 1126–1130, <https://doi.org/10.1093/molbev/msv333>.
- A.V. McDonnell, T. Jiang, A.E. Keating, B. Berger. Paircoil2: Improved prediction of coiled coils from sequence. *Bioinformatics Vol.* 22(3) (2006).
- Aron Marchler-Bauer, Huaiyu Mi, Darren A Natale, Marco Necci, Christine A Orengo, Arun P Pandurangan, Catherine Rivoire, Christian J A Sigrist, Ian Sillitoe, Narmada Thanki, Paul D Thomas, Silvio C E Tosatto, Cathy H Wu, Alex Bateman, Robert D Finn *Nucleic Acids Research* (2020), gkaa977, PMID: [33156333](https://pubmed.ncbi.nlm.nih.gov/33156333/)
- Ashburner et al. Gene ontology: tool for the unification of biology. *Nat Genet.* May 2000;25(1):25-9.
- Ashkenazy H., Abadi S., Martz E., Chay O., Mayrose I., Pupko T., and Ben-Tal N. 2016, ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucl. Acids Res.* 2016; DOI: 10.1093/nar/gkw408; PMID: 27166375 [\[ABS\]](#), [\[PDF\]](#)
- Bálint Mészáros, Gábor Erdős, Zsuzsanna Dosztányi, [IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding](#). *Nucleic Acids Research* 2018;46(W1):W329-W337.
- Bonnie Berger, David B. Wilson, Ethan Wolf, Theodore Tonchev, Mari Milla, and Peter S. Kim. Predicting Coiled Coils by Use of Pairwise Residue Correlations, *Proceedings of the National Academy of Science USA*, vol 92, aug 1995, pp. 8259-8263.
- Buchan DWA, Jones DT (2019). The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/qkz297>
- B. Webb, A. Sali. Comparative Protein Structure Modeling Using Modeller. *Current Protocols in Bioinformatics* 54, John Wiley & Sons, Inc., 5.6.1-5.6.37, 2016.
- Cai, YD., Liu, XJ., Xu, Xb. et al. Support Vector Machines for predicting protein structural class. *BMC Bioinformatics* 2, 3 (2001). <https://doi.org/10.1186/1471-2105-2-3>.
- Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, 50. AmiGO Hub, Web Presence Working Group. AmiGO: online access to ontology and annotation data. *Bioinformatics.* Jan 2009;25(2):288-289.
- CoDNaS 2.0: A comprehensive database of protein conformational diversity in the native state. Alexander M. Monzon; Cristian O. Rohr; María Silvina Fornasari; Gustavo Parisi. *Database.* Oxford Journals. Accepted on March 2016. DOI:10.1093/database/baw038. [Pubmed](#)
- De Castro E, Sigrist CJA, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, Bairoch A, Hulo N, ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res.* 2006 Jul 1;34(Web Server issue):W362-5. PubMed:16845026 [Full text] [PDF version]
- DTU Health Tech, (2017). TMHMM: Prediction of transmembrane helices in proteins, Server v. 2.0. Center for Biological Sequence Analysis, the bioinformatic unit, Technical University of Denmark.
- Elisa Cilia, Rita Pancsa, Peter Tompa, Tom Lenaerts, and Wim Vranken, From protein sequence to dynamics and disorder with DynaMine. *Nature Communications* 4:2741 doi: 10.1038/ncomms3741 (2013).
- Gabler F, Nam SZ, Till S, Mirdita M, Steinegger M, Söding J, Lupas AN, Alva V, Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. *Curr Protoc Bioinformatics.* (2020) Dec;72(1):e108. doi: 10.1002/cpbi.108.
- Greene, L. H., Lewis, T. E., Addou, S., Cuff, A., Dallman, T., Dibley, M., Redfern, O., et al. (2007). The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Research*, 35(Database issue), D291-7. doi:10.1093/nar/gkl959; <https://www.cathdb.info/>.
- Geer LY, Domrachev M, Lipman DJ, Bryant SH. CDART: protein homology by domain architecture. *Genome Res.* 2002 Oct;12(10):1619-23.

Greenwood Tiffany A, Lazzeroni Laura C, Murray Sarah S, Cadenhead Kristin S, Calkins Monica E, Dobie Dorcas J, Green Michael F, Gur Raquel E, Gur Ruben C, Hardiman Gary, Kelsoe John R, Leonard Sherry, Light Gregory A, Nuechterlein Keith H, Olincy Ann, Radant Allen D, Schork Nicholas J, Seidman Larry J, Siever Larry J, Silverman Jeremy M, Stone William S, Swerdlow Neal R, Tsuang Debby W, Tsuang Ming T, Turetsky Bruce I, & Freedman R. & Braff David L (2011). Analysis of 94 candidate genes and 12 endophenotypes for schizophrenia from the Consortium on the Genetics of Schizophrenia. *The American journal of psychiatry*. DOI: [10.1176/appi.ajp.2011.10050723](https://doi.org/10.1176/appi.ajp.2011.10050723)

GRIN3A, Uniprot: the universal protein knowledgebase (2021) . <https://www.uniprot.org/uniprot/Q8TCU5>

Guindon S., Gascuel O. PhyML : "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood." *Systematic Biology*. 2003 52(5):696-704.

Gurillo Muñoz P. (2016). Revisión de los Trastornos del Espectro Psicótico: Sus características genéticas y función de las neurexinas. GRIN Verlag. <https://www.grin.com/document/439363>

Hoffman, K & Stoffel, W (1993). TMbase - A database of membrane spanning proteins segments, *Biol. Chem. Hoppe-Seyler* 374,166.

Kelley, L., Mezulis, S., Yates, C. et al. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* 10, 845–858 (2015). <https://doi.org/10.1038/nprot.2015.053>

Linding R, Russell RB, Neduva V, Gibson TJ. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res*. 2003 Jul 1;31(13):3701-8. doi: 10.1093/nar/gkg519. PMID: 12824398; PMCID: PMC169197.

Lisanna Paladin, Martina Bevilacqua, Sara Errigo, Damiano Piovesan, Ivan Mičetić, Marco Necci, Alexander Miguel Monzon, Maria Laura Fabre, Jose Luis Lopez, Juliet F Nilsson, Javier Rios, Pablo Lorenzano Menna, Maia Cabrera, Martin Gonzalez Buitron, Mariane Gonçalves Kulik, Sebastian Fernandez-Alberti, Maria Silvina Fornasari, Gustavo Parisi, Antonio Lagares, Layla Hirsh, Miguel A Andrade-Navarro, Andrey V Kajava, Silvio C E Tosatto, *RepeatsDB in 2021: improved data and extended classification for protein tandem repeat structures*, *Nucleic Acids Research*, Volume 49, Issue D1, 8 January 2021, Pages D452–D457, <https://doi.org/10.1093/nar/gkaa1097>

Lukas Käll, Anders Krogh and Erik L. L. Sonnhammer. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server, *Nucleic Acids Res.*, 35:W429-32, July 2007 ([doi](https://doi.org/10.1093/nar/gkh107)) ([PubMed](https://pubmed.ncbi.nlm.nih.gov/17511411/))

Lupas, A., Van Dyke, M., and Stock, J. (1991). COILS: Predicting Coiled Coils from Protein Sequences, *Science* 252:1162-1164.

MAPPING and MOLECULAR GENETICS. <http://www.omim.org/entry/181500>

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1). <https://doi.org/10.1093/nar/gkaa913>; <http://pfam.xfam.org/>.

M.Torrisi, M.Kaleel, G.Pollastri. Deeper Profiles and Cascaded Recurrent and Convolutional Neural Networks for state-of-the-art Protein Secondary Structure Prediction, (2019) *Scientific Reports*, 9: 12374, 2019, doi: 10.1038/s41598-019-48786-x

Notredame, Higgins, Heringa, T-Coffee: A novel method for multiple sequence alignments. *JMB*, 302 (205-217) 2000.

Piovesan D, Necci M, Escobedo N, Monzon AM, Hatos A, Mičetić I, Quaglia F, Paladin L, Ramasamy P, Dosztányi Z, Vranken WF, Davey N, Parisi G, Fuxreiter M and Tosatto SCE, (2020), *MobiDB: intrinsically disordered proteins in 2021*. *Nucleic Acid Research*. gkaa1058. [PubMed](https://pubmed.ncbi.nlm.nih.gov/32411111/)

Pond, S. L. K., Frost, S. D. W., & Muse, S. V. (2005). HyPhy : hypothesis testing using phylogenies. *Bioinformatics*, 21(5), 676-679. doi:10.1093/bioinformatics/bti079.

Sánchez de las Matas Martín, María del Carmen. Teoría de la mente y esquizofrenia: aspectos conceptuales y evolutivos. *InterSedes: Revista de las Sedes Regionales*, vol. XV, núm. 30. (2014), pp. 169-196 Universidad de Costa Rica Ciudad Universitaria Carlos Monge Alfaro, Costa Rica.

S.C. Potter, A. Luciani, S.R. Eddy Y. Park, R. Lopez and R.D. Finn. (2018) HMMER web server: *Nucleic Acids Research*. Web Server Issue 46:W200-W204.

Sigrist CJA, de Castro E, Cerutti L, Cuče BA, Hulo N, Bridge A, Bougueleret L, Xenarios I. New and continuing developments at PROSITE. *Nucleic Acids Res*. 2012; doi: 1093/nar/gks1067. [PubMed:23161676](https://pubmed.ncbi.nlm.nih.gov/23161676/) [Full text] [PDF version]

Stephen K Burley, Charmi Bhikadiya, Chunxiao Bi, Sebastian Bittrich, Li Chen, Gregg V Crichlow, Cole H Christie, Kenneth Dalenberg, Luigi Di Costanzo, Jose M Duarte, Shuchismita Dutta, Zukang Feng, Sai Ganesan, David S Goodsell, Sutapa Ghosh, Rachel Kramer Green, Vladimir Guranović, Dmytro Guzenko, Brian P Hudson, Catherine L Lawson, Yuhe Liang, Robert Lowe, Harry Namkoong, Ezra Peisach, Irina Persikova, Chris Randle, Alexander Rose, Yana Rose, Andrej Sali, Joan Segura, Monica Sekharan, Chenghua Shao, Yi-Ping Tao, Maria Voigt, John D Westbrook, Jasmine Y Young, Christine Zardecki, Marina Zhuravleva, RCSB Protein Data Bank:

powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences, *Nucleic Acids Research*, Volume 49, Issue D1, 8 January 2021, Pages D437–D451, <https://doi.org/10.1093/nar/gkaa1038>

Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I. New and continuing developments at PROSITE. *Nucleic Acids Res.* 2012; doi: 1093/nar/gks1067. PubMed:23161676 [Full text] [PDF version]

Stephen K Burley, Charmi Bhikadiya, Chunxiao Bi, Sebastian Bittrich, Li Chen, Gregg V Crichlow, Cole H Christie, Kenneth Dalenberg, Luigi Di Costanzo, Jose M Duarte, Shuchismita Dutta, Zukang Feng, Sai Ganesan, David S Goodsell, Sutapa Ghosh, Rachel Kramer Green, Vladimir Guranović, Dmytro Guzenko, Brian P Hudson, Catherine L Lawson, Yuhe Liang, Robert Lowe, Harry Namkoong, Ezra Peisach, Irina Persikova, Chris Randle, Alexander Rose, Yana Rose, Andrej Sali, Joan Segura, Monica Sekharan, Chenghua Shao, Yi-Ping Tao, Maria Voigt, John D Westbrook, Jasmine Y Young, Christine Zardecki, Marina Zhuravleva, RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences, *Nucleic Acids Research*, Volume 49, Issue D1, 8 January 2021, Pages D437–D451, <https://doi.org/10.1093/nar/gkaa1038>

The InterPro protein families and domains database: 20 years on. Matthias Blum, Hsin-Yu Chang, Sara Chuguransky, Tiago Grego, Swaathi Kandasamy, Alex Mitchell, Gift Nuka, Typhaine Paysan-Lafosse, Matloob Qureshi, Shriya Raj, Lorna Richardson, Gustavo A Salazar, Lowri Williams, Peer Bork, Alan Bridge, Julian Gough, Daniel H Haft, Ivica Letunic, Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, Thanki N, Yamashita RA, Yang M, Zhang D, Zheng C, Lanczycki CJ, Marchler-Bauer A. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* 2020 Jan 8;48(D1):D265-D268. doi: 10.1093/nar/gkz991. (Epub 2019 Nov 28.) [PubMed PMID: 31777944] [Full Text at Oxford Academic]

The Gene Ontology resource: enriching a GOLD mine. *Nucleic Acids Res.* Jan 2021;49(D1):D325-D334.

The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.

The UniProt Consortium, UniProt: the universal protein knowledgebase in 2021, *Nucleic Acids Research*, Volume 49, Issue D1, 8 January 2021, Pages D480–D489, <https://doi.org/10.1093/nar/gkaa1100>

Wiederstein & Sippl (2007), ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research* 35, W407-W410.

Wu S, Zhang Y. LOMETS: A local meta-threading-server for protein structure prediction. *Nucleic Acids Research.* 35, 3375-3382 (2007).

Xu D., Jaroszewski L., Li Z., Godzik A. (2013). FFAS-3D: Improving fold recognition by including optimized structural features and template re-ranking. *Bioinformatics* (2013) doi: 10.1093. PubMed

Zhang M, Liu D, Tang J, Feng Y, Wang T, Dobbin KK, Schliekelman P, Zhao S. SEG - A Software Program for Finding Somatic Copy Number Alterations in Whole Genome Sequencing Data of Cancer. *Comput Struct Biotechnol J.* 2018 Sep 7;16:335-341. eCollection 2018. <https://doi.org/10.1016/j.csbj.2018.09.001>

Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V, A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core.. *J Mol Biol.* 2018 Jul 20. S0022-2836(17)30587-9.

Zhu F, Han LY, Chen X, Lin HH, Ong S, Xie B, Zhang HL, Chen YZ. Homology-free prediction of functional class of proteins and peptides by support vector machines. *Curr Protein Pept Sci.* 2008 Feb;9(1):70-95. doi: [10.2174/138920308783565697](https://doi.org/10.2174/138920308783565697). PMID: 18336324.

Estudios bioquímicos de la ATP citrato liasa de bacterias simbiotes de *Alvinella pompejana* (Anellida: Poliqueta)

Ignacio Pavía

Cátedra de Bioinformática, Área de Biotecnología y Biología Molecular, Departamento de Ciencias Biológicas, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, La Plata, Argentina.

RESUMEN

En el presente trabajo se estudió la ATP citrato liasa de bacterias simbiotes de *Alvinella pompejana*, un poliqueto que habita en chimeneas hidrotermales. Se llevaron a cabo estudios bioinformáticos para determinar sus homólogos, estructura y relaciones filogenéticas. Se determinó que se trata de una proteína formada por dos dominios, uno de unión a citrato y otro con actividad carboxilato-amina ligasa. Posee alto contenido de estructura secundaria, no posee regiones transmembrana ni desordenadas. Se hallaron abundantes homólogos en los diversos taxa de los tres reinos del árbol de la vida. Se demostró que a lo largo de la evolución se conservó el dominio de unión a citrato. El alineamiento múltiple de los homólogos posee información para sustentar varios clados.

PALABRAS CLAVE: ATP citrato liasa, procariontes, bioinformática

INTRODUCCIÓN

La ATP citrato liasa es una enzima clave en el ciclo reductor del ácido tricarbóxico (rTCA) de la fijación de dióxido de carbono durante el crecimiento autótrofo de organismos procariontes. Esta enzima cataliza la escisión del citrato en acetil coenzima A y oxalacetato, en una reacción inversa a la del ciclo de TCA. Una reversión de todo el ciclo genera una molécula de oxalacetato de cuatro moléculas de CO₂.

La mayoría de los quimioautótrofos caracterizados asociados con invertebrados de entornos marinos, incluidos los respiraderos hidrotermales, utilizan la vía Calvin-Benson para la fijación de dióxido de carbono. El ciclo de rTCA se considera una vía alternativa de fijación de carbono y se ha demostrado sólo en unos pocos procariontes.

En el presente trabajo se analizó la ATP citrato liasa de bacterias episimbiotes del gusano poliqueto *Alvinella pompejana*, el cual habita a altas temperaturas en chimeneas hidrotermales de la Dorsal del Pacífico Oriental.

Las caracterizaciones moleculares de la población de simbiotes de *A. pompejana* indican que la mayoría de los simbiotes se agrupan dentro de un solo clado monofilético dentro de la subdivisión épsilon de las Proteobacterias (Bacteria › Proteobacteria › delta/epsilon subdivisions › Epsilonproteobacteria › Campylobacterales › Campylobacteraceae › unclassified Campylobacteraceae)

Los objetivos del presente trabajo fueron analizar la ATP citrato liasa secuencialmente, buscando la presencia de estructuras secundarias, regiones desordenadas, regiones transmembrana y realizar un modelado por homología. Por otro lado, se propuso realizar una búsqueda de secuencias homólogas tanto en taxa cercanos como lejanos filogenéticamente para establecer relaciones y visualizar qué regiones se conservaron a lo largo de la evolución.

MÉTODOS Y RESULTADOS

Predicción de elementos estructurales

Se comenzó con la búsqueda de la proteína en Uniprot (Código: Q6W3M3), se trata de una proteína de 447 aminoácidos.

Para conocer su estructura se realizó una búsqueda de dominios conservados en Pfam (<http://pfam.xfam.org/>), detectándose 2 dominios. El dominio ATP-grasp_2, que va del aminoácido 5 al 228, y se encuentra presente en enzimas que poseen actividad carboxilato-amina ligasa dependiente de ATP, para las cuales es probable que sus mecanismos catalíticos incluyan intermedios de acilfosfato. El otro dominio es el dominio Citrate_bind que comienza en el aminoácido 266 y termina en el 445, dicho dominio posee un sitio de unión a citrato (Fig 1).

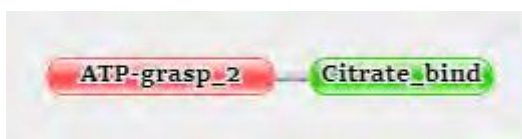


Fig 1. Dominios ATP-grasp_2 y Citrate_bind precedidos por Pfam (<http://pfam.xfam.org/>)

A través del sitio web Jpred (<http://www.compbio.dundee.ac.uk/jpred/>) se determinó la presencia de estructuras secundarias. Los resultados obtenidos demuestran la presencia de α -helices y láminas β -plegada en proporciones similares, siendo las primeras más abundantes (Fig 2).

No se detectaron dominios transmembrana ni regiones desordenadas. Para dichos análisis se utilizó Phobius (<https://phobius.sbc.su.se/>) y MobiDM (<https://mobidb.bio.unipd.it/>).

Búsqueda de homólogos

Se buscaron homólogos mediante BLAST en la página del NCBI. Las búsquedas se realizaron diferencialmente para distintos grupos de organismos (tabla 1) y utilizando parámetros por defecto. Se seleccionó al menos una secuencia de cada grupo, seleccionando un total de 30 secuencias. Además, se realizó un PSIBLAST buscando homólogos remotos encontrándose secuencias con hasta un 34% de identidad siendo el valor más bajo registrado.

Alineamiento múltiple

Se realizó un alineamiento múltiple por Clustal Omega para visualizar el grado de conservación de los 2 dominios encontrados en la proteína estudiada. Se utilizaron los parámetros por defecto. Los resultados del alineamiento mostraron una sola región bien conservada que va desde el aminoácido número 279 al 443. Dicha región coincide con el dominio Citrate_bind (Fig 3).

Modelado por homología

Se realizó un modelado por homología utilizando Modeller. Se comenzó con la búsqueda de un template, para lo cual realizamos un PSI-BLAST en el NCBI contra la PDB, siendo seleccionada la proteína Chain A, Citrate lyase, subunit 1 de *Methanotherx soehngeni* (Archaea; Código de acceso 6HXI_A) la cual posee un 49% de identidad con nuestra proteína. Una vez obtenidos 10 modelos, se evaluaron en ProsaII

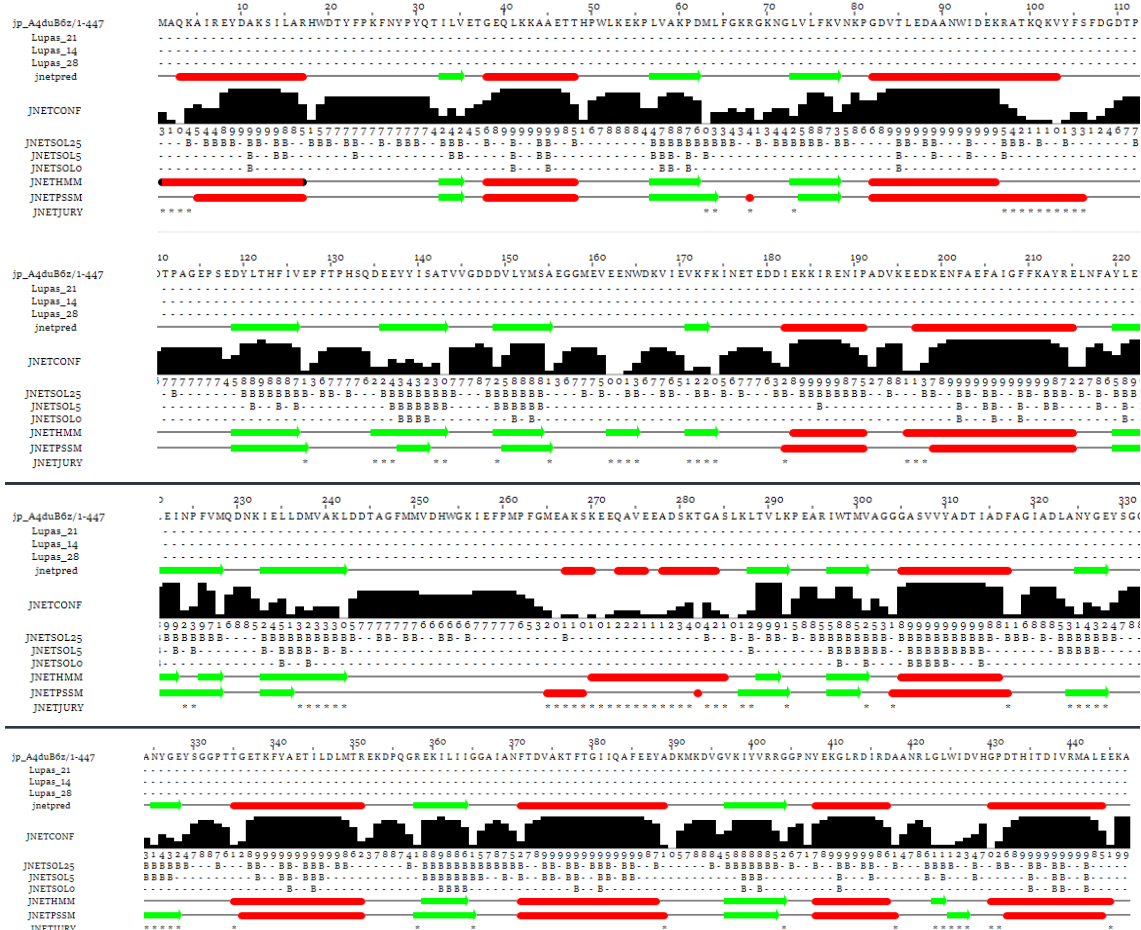


Fig 2. Estimación de estructura secundaria a través de Jpred (<http://www.compbio.dundee.ac.uk/jpred/>). Las líneas rojas y verdes indican estructuras secundarias, α -hélice y lamina β - plegada respectivamente.

Tabla 1. Proteínas de diferentes grupos de organismos seleccionados a partir del BLAST.

Grupo taxonómico	Especie	Código NCBI
Chlorophyta (alga verde)	<i>Micractinium conductrix</i>	PSC76528.1
Annelida	<i>Dimorphilus gyrociolatus</i>	CAD5123422.1
Annelida	<i>Capitella teleta</i>	ELU08772.1
Amphibia	<i>Xenopus tropicalis</i>	NP_001008028.1
Embryophyta (Plantas terrestres)	<i>Citrus clementina</i>	XP_006438324.1
Embryophyta	<i>Populus alba</i>	XP_034899784.1
Archaea	<i>Candidatus Methanoperedenaceae</i>	CAD7776443.1
Archaea	<i>Thermoplasmatales archaeon</i>	HDM67402.1
Aves	<i>Lepidothrix coronata</i>	XP_017689586.1
Aves	<i>Anhinga rufa</i>	NXT90945.1
Bryozoa	<i>Bugula neritina</i>	KAF6018373.1
Cephalochordata	<i>Branchiostoma belcheri</i>	XP_019635091.1
Arachnida	<i>Tropilaelaps mercedesae</i>	OQR79781.1
Chondrichthyes	<i>Chiloscyllium punctatum</i>	GCC36100.1
Crustacea	<i>Tigriopus kingsejongensis</i>	AUZ82868.1
Bacillariophyta (Diatomea)	<i>Fistulifera solaris</i>	GAX17489.1
Echinodermata	<i>Acanthaster planci</i>	XP_022096812.1
Porifera	<i>Amphimedon queenslandica</i>	XP_003386132.1
Fungi	<i>Podila clonocystis</i>	KAG0018024.1
Fungi	<i>Taphrina deformans</i> PYCC 5710	CCG84442.1
Hemichordata	<i>Saccoglossus kowalevskii</i>	XP_006821438.1
Insecta	<i>Oryctes borbonicus</i>	KRT85288.1
Mammalia	<i>Homo Sapiens</i>	3MWE_A
Mammalia	<i>Mus musculus</i>	BAE36874.1
Mollusca	<i>Biomphalaria glabrata</i>	XP_013071117.1
Nematoda	<i>Caenorhabditis brenneri</i>	EGT55543.1
Testudinata	<i>Pelodiscus sinensis</i>	XP_006131005.1
Teleostei	<i>Perca flavescens</i>	TDG97951.1
Rotifera	<i>Brachionus plicatilis</i>	RMZ99671.1
Tardigrada	<i>Ramazzottius varieornatus</i>	GAV06129.1

(<https://prosa.services.came.sbg.ac.at/prosa.php>) seleccionándose el de menor energía global (Fig 4. A). Al encontrarse nuestro modelo dentro del área sombreada se puede establecer que es un buen modelo, ya que para generar dicha distribución se utiliza una colección curada de proteínas de estructura conocida y con buena resolución. Además, se analizó el gráfico de energías locales mostrando mayormente zonas de baja energía (Fig 4. B). El modelo seleccionado se visualizó mediante PyMol (Fig 5).



Fig 3. Fragmento del alineamiento múltiple De la ATP citrato liasa de bacterias episimbiontes del gusano poliqueto *Alvinella pompejana* y secuencias de diversos organismos que se muestran en la tabla 1. Los residuos idénticos conservados en todas las secuencias se encuentran sombreados en negro; los conservados en la mayoría de las secuencias están sombreados en gris oscuro; y los cambios conservadores están en gris claro. Las zonas encuadradas en rojo son las más conservadas.

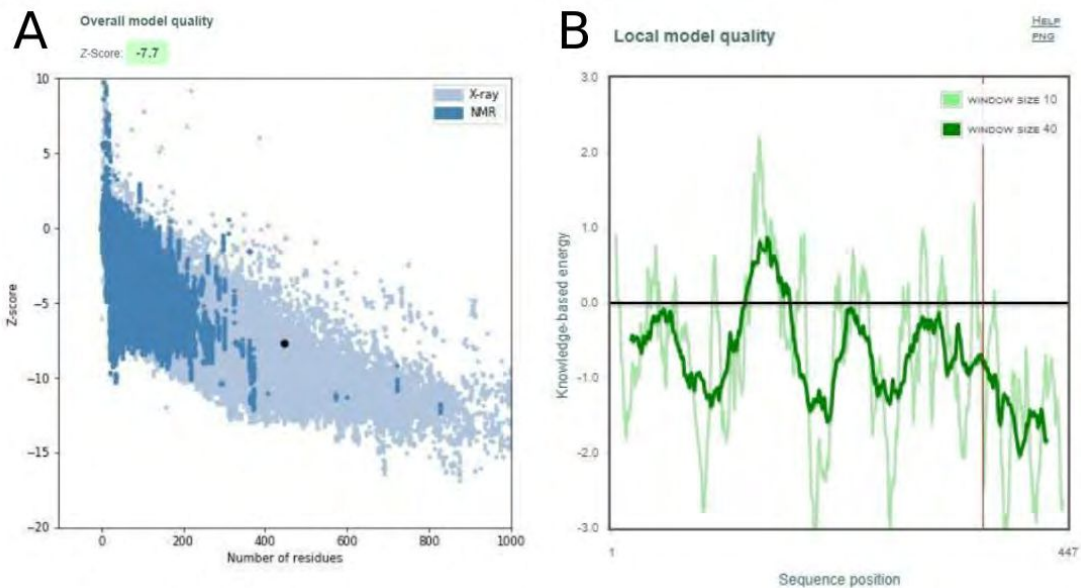


Fig 4. A. Gráfico de distribución de energías en función de la longitud de la estructura, generado con proteínas curadas y de buena resolución. El punto negro representa a nuestro modelo. B. Gráfico de energía local. Muestra la energía de nuestro modelo en función de la posición de los aminoácidos en nuestra secuencia. Se puede observar mayormente zonas de baja energía a excepción de un pico en la posición 150 aproximadamente.

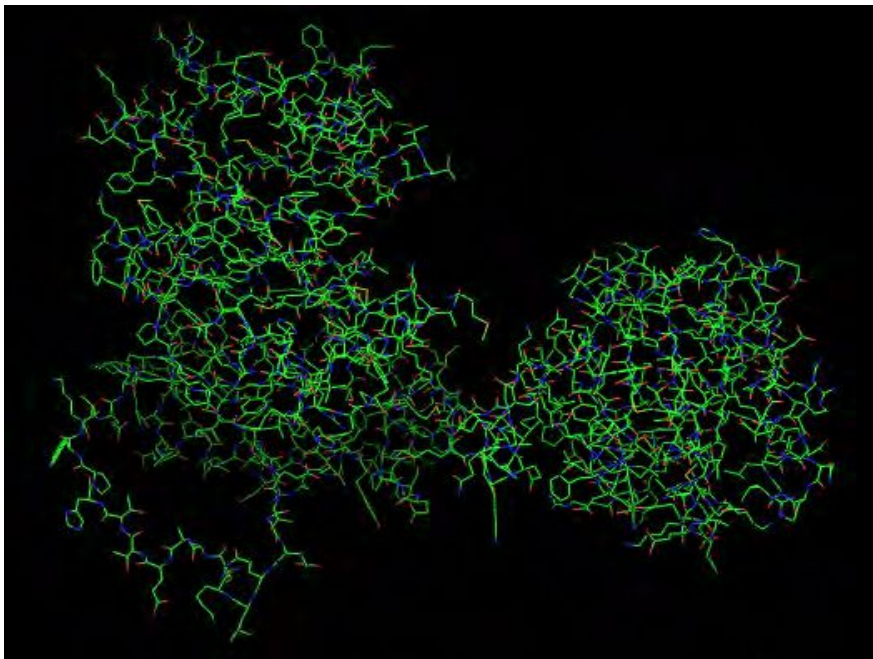


Fig 5. Modelo generado por modelado por homología a través del Modeller.

Inferencia filogenética

El análisis filogenético se llevó a cabo por el método de Maximun Likelihood en el programa PhyML. Para ello se utilizó el alineamiento múltiple realizado anteriormente, un árbol input realizado por Neibor Joining y el modelo evolutivo WAG (Seleccionado como mejor modelo utilizando modeltest en el programa Hyphy). La confianza de las topologías se estimó mediante un bootstraping no paramétrico con 1000 réplicas. Una vez obtenido el árbol se visualizó y editó en el programa FigTree (Fig 6). Se pudo observar que nuestro alineamiento posee mucha información para soportar dos grandes grupos que divergen al principio del árbol, Un primer grupo formado por Bacterias, Arqueas, algas y Plantas terrestres, en el cual están bien soportados todos los nodos entre los diferentes organismos, y un segundo grupo formado por Diatomeas,

Hongos y Animales. En este último grupo hay un subgrupo muy bien soportado el de los Vertebrados (Vertebrata) comenzando con los Chondrictios.

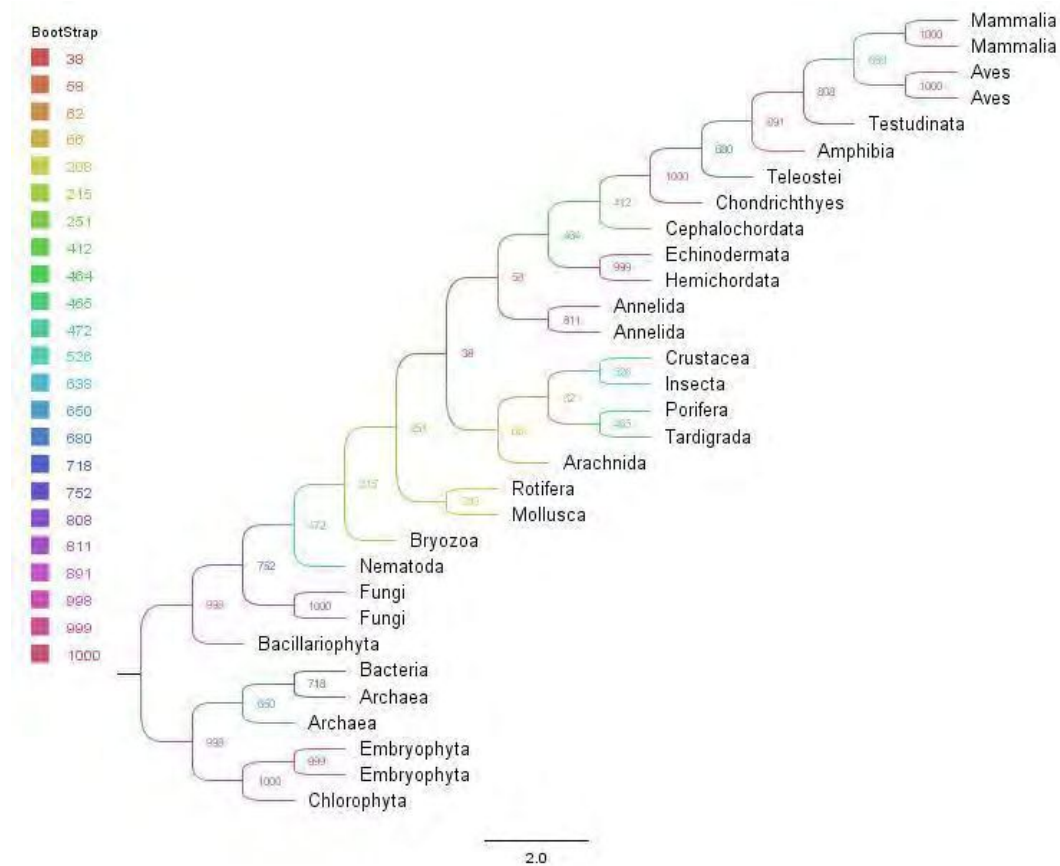


Fig 6. Árbol filogenético de diferentes taxos de organismos (mencionados en la tabla 1).

CONCLUSIONES Y DISCUSIÓN

La proteína Q6W3M3 se trata de una proteína sin regiones transmembrana, que cuenta con alto contenido de estructura secundaria, tanto alfa-hélices como lámina beta-plegada y no posee regiones desordenadas.

Dicha proteína cuenta con gran cantidad de homólogos en diferentes taxones de los 3 dominios (Archaea, Bacteria y Eucharia). Y a lo largo de la evolución pareciese haberse conservado el dominio Citrate_bind.

El análisis filogenético posee información para soportar la relación de varios taxones, incluso en grupos alejados filogenéticamente. En grupos cercanos evolutivamente se puede observar que no ha variado mucho la proteína, ya que, en los grupos duplicados de un mismo taxon siempre se dieron juntos y con un alto valor de bootstrapping.

BIBLIOGRAFÍA

Campbell, B. J., Stein, J. L., Cary, C. S. Evidence of Chemolithoautotrophy in the Bacterial Community Associated with *Alvinella pompejana*, a Hydrothermal Vent Polychaete. 2003. *Applied and environmental microbiology*, vol 69, N° 9, p. 5070–5078.

Di Meo-Savoie, C. A., Luther, G. W., Cary, C. S. Physicochemical characterization of the microhabitat of the epibionts associated with *Alvinella pompejana*, a hydrothermal vent annelid. 2004. *Geochimica et Cosmochimica Acta*, Vol. 68, No. 9, pp. 2055–2066.

Wahlund, T. M., Tabita, F. R., The Reductive Tricarboxylic Acid Cycle of Carbon Dioxide Assimilation: Initial Studies and Purification of ATP-Citrate Lyase from the Green Sulfur Bacterium *Chlorobium tepidum*. 1997. *Journal of Bacteriology*, Vol. 179, No. 15, p. 4859–4867.

Recorrido por Hv1: un canal selectivo de protones

Victoria Pinto

Cátedra de Bioinformática, Área de Biotecnología y Biología Molecular, Departamento de Ciencias Biológicas, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, La Plata, Argentina.

RESUMEN

A lo largo del cuatrimestre se fueron introduciendo diferentes métodos y técnicas bioinformáticas para el estudio de proteínas. Este análisis se propone estudiar de forma secuencial, estructural y funcional al canal voltaje operado en humanos, Hv1.

Para esto, se realizó una búsqueda de homología, fijando los parámetros necesarios. Luego, con alguna de estas secuencias encontradas, se hizo un alineamiento múltiple para corroborar la importancia de Asp112 en la selectividad de los protones.

También se realizó la predicción de estructura secundaria y se modeló la estructura terciaria. Se construyó, además, un árbol filogenético utilizando secuencias de proteínas homólogas de distintos mamíferos que expresan el canal. Por último, se hizo un análisis de la función del canal.

PALABRAS CLAVE: canal de protones voltaje operado; mamíferos; bioinformática

INTRODUCCIÓN

En este trabajo presento un estudio bioinformático sobre la proteína Hv1, un canal de protones voltaje operado en humanos. El gen HVCN1 que codifica para la proteína, se encuentra en el cromosoma 12 de humanos y se expresa en varios tejidos del organismo. Existen cuatro isoformas, de las cuales se trabajará con la isoforma 1, cuya longitud es de 273 aminoácidos (Uniprot ID: Q96D96).

En condiciones fisiológicas normales el pH extracelular es ligeramente alcalino (7,3-7,4); y el pH citosólico, menor que dicho valor (7,2). Dos propiedades celulares dan base a la tendencia del citosol a acidificarse: en primer lugar el potencial de membrana con el interior celular negativo respecto al extracelular, brinda una fuerza impulsora importante para la entrada de protones y la salida de bases negativamente cargadas, como el HCO_3^- . En segundo lugar, diversas reacciones metabólicas vitales para la célula, como la producción de ATP por la glicólisis en el citoplasma o la fosforilación oxidativa en la mitocondria, generan H^+ como productos de la reacción, que aumentan cuando se incrementa la actividad celular. La acumulación gradual de H^+ es compensada mediante su extrusión continua a través de la membrana plasmática por distintos mecanismos.

El canal Hv1 se activa con la despolarización de la membrana celular, mediando una corriente saliente de cargas positivas que tiende a hiperpolarizar la membrana. Además, se encuentra en su conformación abierta sólo cuando el gradiente electroquímico es favorable a la salida espontánea de protones, por lo cual se puede considerar al canal como un estricto extrusor pasivo de H^+ . Esta fuerte dependencia con el pH indica que su actividad será mayor al aumentar el pH extracelular o disminuir el pH intracelular.

Con respecto a la inactivación, evidencias experimentales indican que los canales Hv1 no presentan en esta situación una dependencia con el voltaje. Sin embargo, cuando se evocan corrientes de H^+ demasiado prolongadas (~20 seg.) Se puede observar un decaimiento de la corriente que se asemeja al proceso de

inactivación. Se vio que la misma corriente mantenida en el tiempo es la que produce un cambio en la concentración de H⁺, modificando su gradiente. Esto reduce la fuerza impulsora, produciendo una disminución de la magnitud de la corriente cada vez mayor, de forma similar a lo que ocurre cuando los canales presentan un estado inactivo.

En contraste con los típicos canales de cationes dependientes de voltaje, Hv1 contiene solo el dominio de detección de voltaje (VSD) y carece del dominio de poro central (esencial para otros canales de cationes). La secuencia de aminoácidos de Hv1 está relacionada con el VSD de otros canales de cationes. Otra diferencia es que Hv1 es un dímero, donde cada subunidad contiene tanto el sensor de voltaje como la vía de los protones. Se requieren regiones N- y C-terminales del canal para la dimerización, logrando una activación cooperativa (figura 1). Evidencias muestran que el canal Hv1 también podría ser funcional como monómero. Se postula que los dímeros se mantienen por interacciones Coiled-Coil entre los dominios intracelulares C-terminal, por lo que si se remueve o mutan los dominios C-terminal, podría obtenerse la expresión funcional del monómero (Asuaje, 2018).

Los canales Hv1 ofrecen a la célula que lo expresa, una vía de pasaje de H⁺ que permite la regulación del pH_i sin gasto de energía metabólica. En este sentido, se ha comprobado que la función del canal es muy importante en la recuperación del pH_i de diversos tipos celulares luego de ser sometidos a cargas ácidas: neutrófilos y basófilos humanos.

Además, está involucrado en enfermedades como el cáncer (Bayrhuber et al, 2019):

-La sobreexpresión de Hv1 se correlaciona significativamente con el tamaño del tumor en el cáncer de mama.

-Una isoforma más corta de Hv1 se enriquece específicamente en células de leucemia linfocítica crónica, lo que hace que Hv1 sea un objetivo farmacológico potencial.

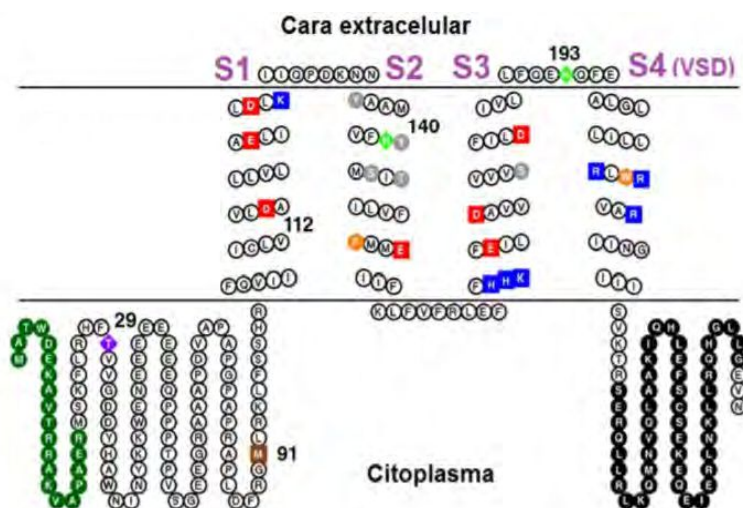


FIGURA 1. Esquema de Hv1 con sus aminoácidos. Se marcan algunos aminoácidos importantes, estudiados previamente, como el caso del Aspartato en la posición 112 (Asuaje, 2018).

MÉTODOS Y RESULTADOS

De la base de datos Uniprot se obtuvo la secuencia en formato FASTA, la cual se utilizó para la búsqueda de homólogos en BLAST. Como parámetros para asegurar homología, se aceptaron aquellos que cumplan con: un coverage del 70% (cuánto cubre la secuencia de la base de datos a la secuencia

query), un porcentaje de identidad de al menos 25% (ya que se trata de una proteína de casi 273 aminoácidos), con E-value cuyo valor máximo sea 10^{-6} . El número de hits seleccionado fue de 5000 y el word size fue modificado a 6 para una mayor sensibilidad. En la corrida del BLAST se obtuvieron 973 secuencias dentro de los parámetros mencionados anteriormente. Se encontraron principalmente en organismos eucariotas, con predominio en animales, y en mucho menor proporción en bacterias.

Adicionalmente se realizó una búsqueda por PSIBLAST con el objetivo de encontrar homólogos más remotos. Se corrió PSIBLAST con un número de hits de 500 y luego de 3 iteraciones, se obtuvieron 500 secuencias con porcentajes de identidad mayores al 30%, coverage del 70%, y E-value bajos. En especial, pudieron encontrarse homólogos remotos que en una corrida de BLAST no llegaban a encontrarse.

Del total de las secuencias homólogas encontradas usando BLAST, se seleccionaron 12, cada una perteneciente a una especie distinta y se realizó un alineamiento múltiple utilizando T-coffee (Fig.2). El aminoácido Asp112 es crucial para la selectividad de los protones. Se muestra en la figura parte del alineamiento en donde se observa que el Asp112 se encuentra 100% conservado.

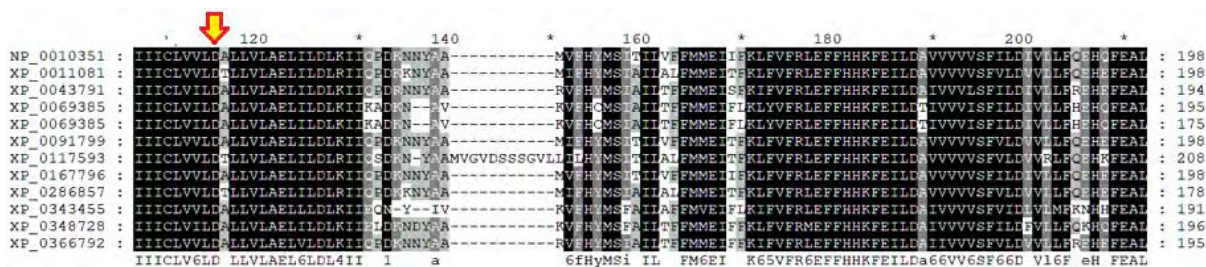


FIGURA 2. Visualización de una sección del alineamiento múltiple de doce secuencias homólogas.

Por otro lado, se realizó una búsqueda con la secuencia query en Pfam, una base de datos HMM de dominios secuenciales que pueden o no coincidir con dominios estructurales. Los resultados permitieron reconocer un dominio de transporte de iones, una región C-terminal del canal Hv1, dos regiones coiled-coil, cuatro sitios transmembrana y una zona desordenada (Fig. 3).



FIGURA 3. Esquema de los resultados obtenidos en Pfam.

El dominio ion_trans hace referencia a un dominio encontrado en canales de sodio, potasio y calcio. Esta familia tiene seis hélices transmembrana, donde las últimas dos flanquean un loop encargado de la selectividad iónica. En el caso de Hv1, no comparte la totalidad del dominio ya que sólo posee cuatro hélices transmembrana.

La familia de este dominio está formada por 3908 especies distintas, distribuidas principalmente en eucariotas y bacterias, y en menor medida en arqueobacterias (Fig. A1).

Con respecto al segundo dominio, VGPC1_C, hace referencia a la región C-terminal del canal de protones voltaje operado encontrado en eucariotas, necesario para que se dé la dimerización, ya que esta se mantiene por interacciones de hélices superenrolladas entre los dominios intracelulares C terminal.

Además, la región es esencial para la localización de la proteína en una membrana intracelular. La familia de este dominio está formada por 134 especies distintas, del dominio eucariota (Fig. A2).

A continuación, se realizó la predicción de estructura secundaria utilizando el programa Quick2D (Gabler, et al. 2020) (Fig. 4):

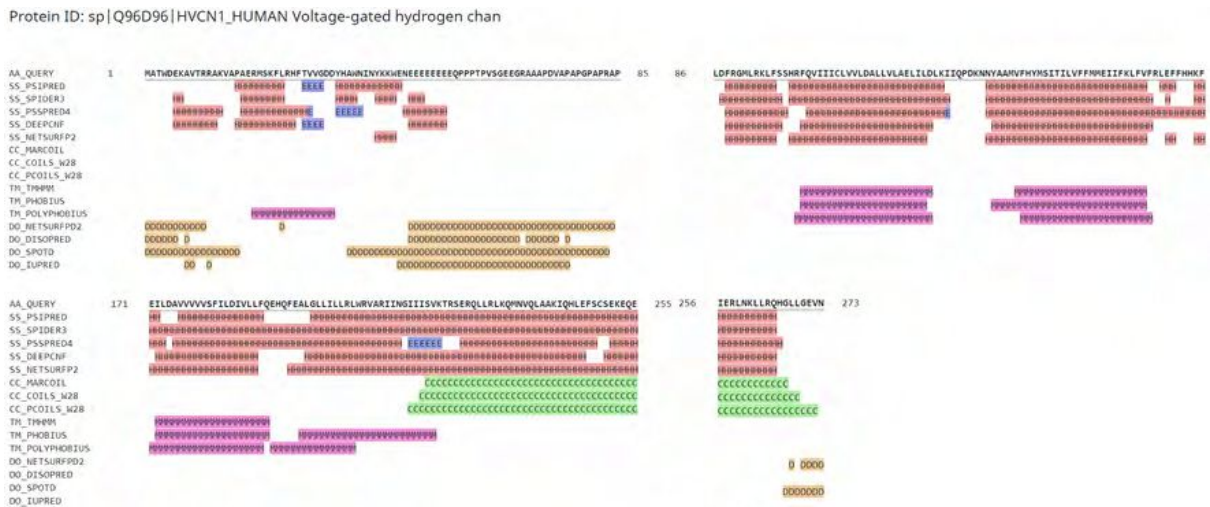


FIGURA 4. Predicción de la estructura secundaria de Hv1. Se representa en rosa las regiones α -hélice, regiones coiled-coil en verde, regiones transmembrana en violeta y las desordenadas en marrón.

Se puede ver mayormente regiones α -hélice, coiled-coil y transmembrana. Se visualiza, además, una pequeña porción desordenada.

Para profundizar en cada una de las estructuras y corroborar los resultados anteriores, se realizó una predicción con distintos servidores de regiones desordenadas, transmembrana y coiled-coil. Para la predicción de desorden se obtuvo la presencia de éste en una pequeña porción de la proteína, utilizando <https://mobidb.org/> (Fig. 5).

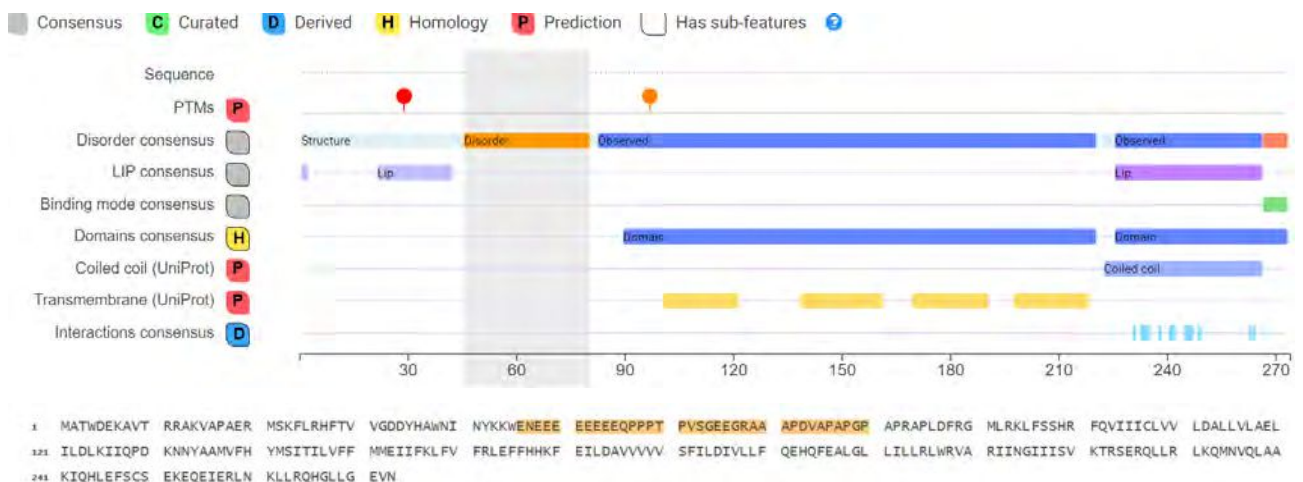


FIGURA 5. Resultado de la predicción de desorden utilizando <https://mobidb.org/>

Para la predicción de regiones transmembrana, se utilizó el programa TMHMM (Fig.A3), mediante el cual se ve una probabilidad alta (cercana a 1) de encontrar tres segmentos transmembrana, mientras que el cuarto segmento tiene una probabilidad baja, de 0,3. Sin embargo Quick2D y Phobius (Käll et al., 2004), encuentran altas posibilidades de existencia de cuatro de estas regiones (Fig.6):

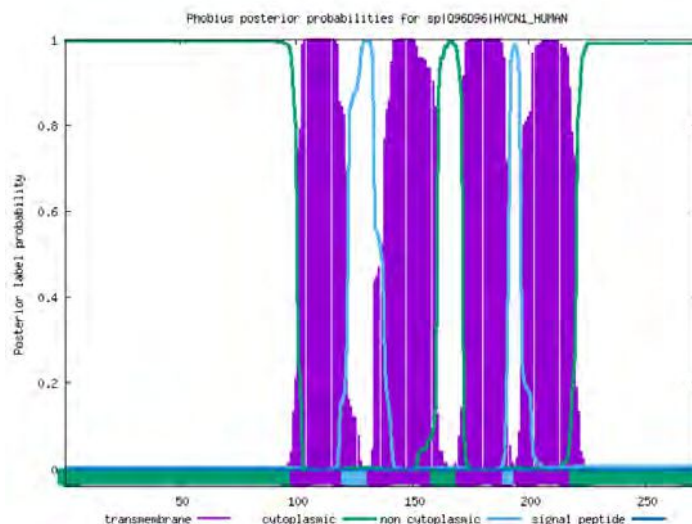


FIGURA 6. Predicción de regiones transmembrana con Phobius. Se ven cuatro regiones transmembrana: del aminoácido 101 al 122, del aminoácido 134 al 160, del aminoácido 172 al 191, y del aminoácido 197 al 220.

Por otro lado, para el análisis de regiones coiled-coil se utilizó el programa Coils (Fig.7). Predice, usando la ventana más grande, una región coiled-coil cercana al extremo C-terminal (coincidente con la predicción realizada con Quick2D). Mientras que con ventanas más chicas, existe probabilidad de encontrar además una región adicional en la zona media de la proteína.

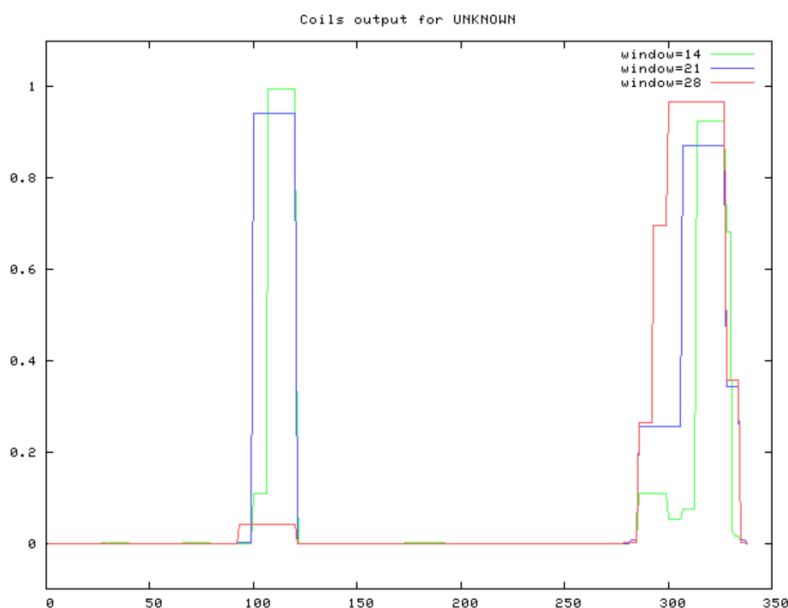


FIGURA 7. Predicción de regiones coiled-coil con Embnet.

Luego de hacer el análisis de la estructura secundaria de la proteína, se realizó un modelado de la estructura terciaria de la misma, usando la proteína homóloga en ratón como template y el programa Modeller. De los 10 modelos obtenidos, todos fueron parecidos y no muy buenos. Esto probablemente se deba a que sólo se encontró una única estructura homóloga (en una única conformación), con una resolución de 3.75 Å. Además, hubo varios residuos perdidos sin coordenadas, por lo que el programa tuvo que trabajar con poca información valiosa. Se utilizó el programa Pymol para la visualización y alineamiento de los modelos con el template (Fig.8).

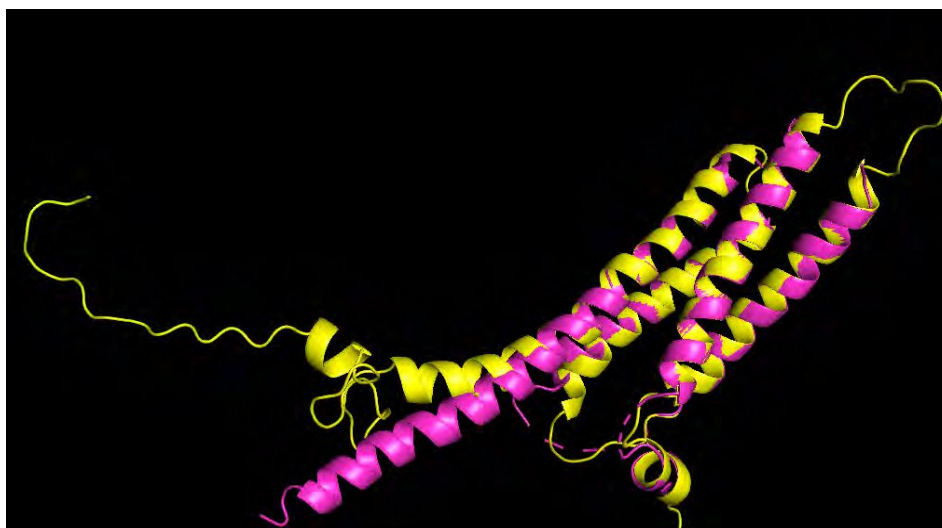


FIGURA 8. Modelado por homología. En rosa el template 3WKV, en amarillo uno de los modelos elegidos. RMSD = 0,243.

La calidad de los modelos obtenidos se analizó con Procheck y Verify (Laskowski et al., 1993; Bowie et al, 1991). El programa Verify analiza el ambiente químico en el cual se encuentra un aminoácido en una posición específica. El resultado arroja que 1,1% de los residuos tienen un score mayor a 0,2. Idealmente, al menos el 80% de los aminoácidos deberían tener ese score (Fig. A4).

Procheck, nos da información sobre el Ramachandran Plot, como forma de evaluar la estereoquímica de los aminoácidos. Es decir, los ángulos psi y phi de cada uno de los aminoácidos de la proteína. Dentro del Plot hay regiones permitidas y no permitidas, por lo que para ser considerada una estructura de buena calidad, al menos el 90% de los aminoácidos debe caer en posiciones permitidas. En este caso, hay un 88,85% dentro de la zona permitida. Lo anterior es otra verificación de que el modelado no fue bueno, tanto en términos de ambiente químico como en aspectos estereoquímicos.

A continuación se realizó un análisis filogenético. Para ello se corrió un BLAST en Uniprot, eligiendo 13 secuencias homólogas, de diferentes organismos mamíferos, ya que el canal sólo se encuentra en éstos. Se realizó un alineamiento múltiple con T-coffee y el análisis filogenético por el método de maximum likelihood utilizando IQ-TREE (incluye la elección del modelo evolutivo, siendo el mejor JTT+G4). El árbol filogenético obtenido fue editado con SNAD y visualizado mediante FigTree (Fig.9). Para el root se usó la opción de punto medio y se utilizó el mismo color para aquellos organismos que compartían el último nodo, es decir, el último evento de especiación en cada caso.

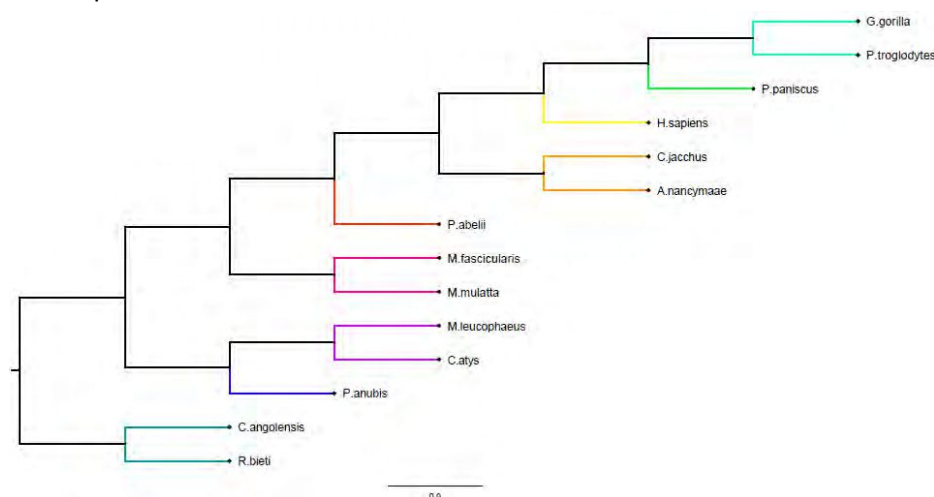


FIGURA 9. Árbol filogenético de la proteína estudiada, obtenido en FigTree

Para terminar de analizar a la proteína, se utilizaron los servidores ConSurf y Evolutionary Trace (Lua et al., 2016). Para el primero, se usó uno de los modelos estructurales obtenidos anteriormente, dando como resultado que 179 de 273 residuos tienen puntuaciones de conservación poco confiables debido a datos insuficientes en el alineamiento múltiple (Fig. 10).



FIGURA 10. Visualización de la conservación de los residuos. El color amarillo evidencia que no hay información suficiente. Desde el celeste al púrpura la conservación aumenta.

El Evolutionary Trace calcula distancias evolutivas para ver el rango relativo de importancia estructural y funcional entre las posiciones de las secuencias alineadas de los homólogos. El rango es menor si las posiciones de la secuencia varían entre homólogos evolutivamente más cercanos y mayor si las posiciones varían entre homólogos evolutivamente distantes.

Se le pidió al programa que haga un alineamiento múltiple con secuencias homólogas seleccionadas por el mismo, arrojando un gráfico Logo y una representación de la secuencia (Fig. 11).

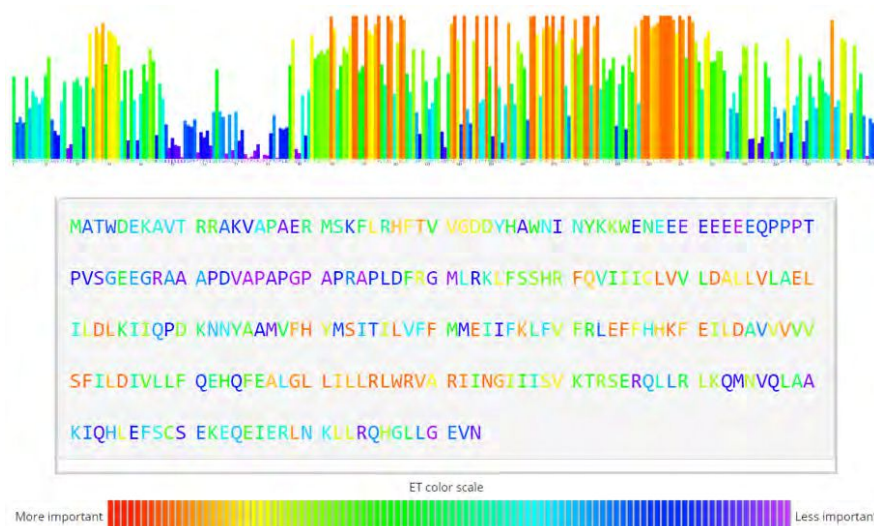


FIGURA 11. Gráfico Logo obtenido con Evolutionary Trace y secuencia en colores del canal.

Se destaca una mayor conservación de la región transmembrana del canal, que coincide con la zona de selectividad iónica. Además del Asp112, existen otros sitios destacados como: Thr29 es el sitio de fosforilación por la protein quinasa C (PKC) lo cual deriva en la potenciación del canal, His140 y His193

coordinan la unión del zinc, el cual ejerce un efecto inhibitorio. (http://sedici.unlp.edu.ar/bitstream/handle/10915/68542/Documento_completo.pdf-PDFA.pdf?sequence=1&isAllowed=y).

CONCLUSIONES Y DISCUSIÓN

Con la información disponible del canal de protones voltaje operado Hv1 se pudo realizar un estudio secuencial, estructural y funcional del mismo, mediante el uso de diferentes programas y servidores.

Se realizó un recorrido por diferentes bases de datos, extrayendo información de las mismas, incluida la secuencia del gen HVCN1 que utilizamos en la mayoría de las instancias.

Se pudo analizar la estructura secundaria del canal y aproximarse a la estructura terciaria usando el mejor homólogo encontrado.

Por otro lado, con algunos de los homólogos cercanos y remotos seleccionados, se pudo construir un árbol filogenético, y conocer, además, posiciones conservadas que podrían llegar a estar relacionadas con la función de la proteína.

BIBLIOGRAFÍA

Asuaje, A. (2018). *Rol del canal de protones operado por voltaje (HVCN1) en la homeostasis ácido-base de células tumorales T humanas y su impacto en la apoptosis (Doctoral dissertation, Universidad Nacional de La Plata)*.

Bayrhuber, M., Maslennikov, I., Kwiatkowski, W., Sobol, A., Wierschem, C., Eichmann, C., Frey, L. & Riek, R. (2019). Nuclear magnetic resonance solution structure and functional behavior of the human proton channel. *Biochemistry*, 58(39), 4017-4027.

Gabler, F., Nam, S. Z., Till, S., Mirdita, M., Steinegger, M., Söding, J., Lupas, A. N. & Alva, V. (2020). Protein sequence analysis using the MPI bioinformatics toolkit. *Current Protocols in Bioinformatics*, 72(1), e108.

TMHMM (Server v. 2.0), <http://www.cbs.dtu.dk/services/TMHMM/>

Käll, L., Krogh, A., & Sonnhammer, E. L. (2004). A combined transmembrane topology and signal peptide prediction method. *Journal of molecular biology*, 338(5), 1027-1036.

Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of applied crystallography*, 26(2), 283-291.

Bowie, J. U., Lüthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253(5016), 164-170.

Lua, R. C., Wilson, S. J., Konecki, D. M., Wilkins, A. D., Venner, E., Morgan, D. H., & Lichtarge, O. (2016). UET: a database of evolutionarily-predicted functional determinants of protein sequences that cluster as functional sites in protein structures. *Nucleic acids research*, 44(D1), D308-D312.

Análisis bioinformático de la enzima Timidilato Sintasa de *Candidatus Poseidoniales archaeon*

Bernarda Pschunder

Cátedra de Bioinformática, Área de Biotecnología y Biología Molecular, Departamento de Ciencias Biológicas, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, La Plata, Argentina.

RESUMEN

En el presente trabajo se genera un modelo estructural tridimensional de la Timidilato Sintasa de *Candidatus Poseidoniales archaeon* utilizando herramientas bioinformáticas de carácter secuencial, estructural, evolutivo y funcional. El modelo encontrado demuestra tener similitud secuencial y estructural con Timidilato Sintasa provenientes de distintos organismos.

Palabras clave: timidilato sintasa, Archea, *Candidatus Poseidoniales*, modelado, bioinformática.

INTRODUCCIÓN

Las arqueas del Grupo Marino II (MGII) representan el grupo de arqueas planctónicas más abundante de las aguas superficiales del océano pero su descripción se encuentra limitada debido a la falta de organismos cultivables representativos y la falta de secuencias genómicas disponibles. A partir del análisis filogenético comparativo de 270 secuencias se propuso que el MGII es un linaje a nivel de orden para el cual se propuso el nombre *Candidatus Poseidoniales* y que comprende a las familias *Candidatus Poseidonaceae* y *Candidatus Thalassarchaeaceae*, dentro de las cuales se pueden distinguir 21 géneros (Rinke et al. 2019).

La proteína que se analizó en este trabajo (UniProt ID: A0A6V8D8A1) corresponde a la enzima Timidilato Sintasa perteneciente a la especie *Candidatus Poseidoniales archaeon* que aún no se clasifica dentro de ninguna de las dos familias mencionadas anteriormente. Es una enzima metiltransferasa cuya función se encuentra relacionada a la biosíntesis de nucleótidos y específicamente cataliza la reacción de 5,10-metileno-tetrahidrofolato con dUMP para dar 7, 8-dihidrofolato y timidilato. Esta enzima se caracteriza por presentar un único dominio, con una longitud de 264 aminoácidos, de la cual se conoce su secuencia y homólogos con 90% y 50% de identidad como lo muestra la base de datos de UniProt. Por el contrario, no se tiene información sobre su estructura secundaria ni tampoco un modelo tridimensional de la proteína.

A partir de esta breve descripción nos propusimos como objetivos principales confirmar lo que se conoce hasta el momento sobre esta enzima y posteriormente realizar un modelado tridimensional con la intención de inferir los posibles sitios de mayor importancia biológica. Para el cumplimiento de estos objetivos se utilizarán herramientas bioinformáticas de carácter secuencial, estructural y evolutivo

MÉTODOS Y RESULTADOS

Búsqueda de homólogos cercanos y remotos utilizando diferentes metodologías

Para el análisis secuencial se utilizan 3 herramientas con fundamentos distintos: en primer instancia se realiza un BLAST, donde la metodología se basa en un análisis secuencia-secuencia, luego se realiza un PsiBlast siguiendo un análisis de tipo secuencia-profile y por último a través de la herramienta HMMER se obtienen resultados con base secuencia-HMM. A continuación, se detallan los parámetros utilizados en cada programa y en la *Tabla 1* se muestra el resumen de los datos más relevantes obtenidos utilizando las diferentes herramientas.

- *Alineamiento secuencia- secuencia con BLAST* utilizando <https://blast.ncbi.nlm.nih.gov/>

Número de secuencias: 1000
 Base de datos: non-redundant protein sequences (nr)
 Umbral de E-value: 0.05
 Tamaño de palabra: 6

- *Alineamiento secuencia-profile con PSI-BLAST:* <https://toolkit.tuebingen.mpg.de/>

Número de secuencias: 1000
 Base de datos: nr50_12_Apr (Default)
 Valor de corte de E-value para inclusión en el profile: 1E-6

- *Alineamiento secuencia-HMM utilizando Hmmer:* <https://www.ebi.ac.uk/Tools/hmmer>

Metodología	BLAST		PsiBlast		HMMER		
	Homólogo más cercano	Homólogo más lejano	Homólogo más lejano (IT1)	Homólogo más lejano (IT2)	Homólogo más lejano (IT1)	Homólogo más lejano (IT2)	Homólogo más lejano (IT3)
%Identidad	100	67,8	15	14	26,9	23,7	16,9
%Similitud	100	86	35	36	77	63,4	57,1
%Gaps	0	0	-	-	-	-	-
E-value	0	3,00E-144	3,00E-08	3,00E-06	1,00E-02	1,00E-02	1,10E-02
Score	555	414	63,2	54	26,6	26,4	26,3

Tabla 1. Resultados más relevantes obtenidos en las búsquedas realizadas en BLAST, PsiBlast y HMMER

En primer lugar, mediante el uso de BLAST se encontró que la secuencia con 100% de identidad y 0% de Gaps corresponde a la secuencia de la Timidilato Sintasa de *Candidatus Poseidoniales archaeon*, información que se corresponde con lo encontrado en la base de datos de UniProt. En segundo lugar, si bien el BLAST es un buen comienzo para encontrar los homólogos cercanos a nuestra secuencia de interés, los homólogos más lejanos tienen un %Identidad del 67% con buenos scores y E-values bajos, por lo tanto con esta herramienta y con la limitante de que sólo se pueden analizar 5000 secuencias, se complica la búsqueda de homólogos más lejanos. Para cumplir este objetivo es necesario utilizar otros algoritmos y otras bases de datos.

Mediante el uso de PsiBlast o HMMER encontramos secuencias con un porcentaje de identidad mucho menor y con E-values mayores. En estos casos donde los %Identidad se encuentran por debajo del 20-30%, no se puede deducir a simple vista si las secuencias alineadas son homólogas o no, por lo que sería conveniente realizar otro tipo de estudios como un estudio estructural.

Por otra parte, la herramienta HMMER nos da como resultado una distribución taxonómica de los hits como se muestra en la Figura 1. Esta representación nos muestra la distribución de los hits donde como se puede observar, la mayor proporción corresponde al dominio Bacteria, siguiendo con el dominio Eukaryota y por último el dominio Archaea. Sabemos que la secuencia analizada corresponde a un organismo del dominio Archaea y el bajo número de hits puede corresponder no a diferencias secuenciales, sino que tal vez las bases de datos no contienen tanta información secuencial proveniente de arqueas. Por el contrario, sí se tiene información del dominio Bacteria y Eukarota y por eso el número de hits podría ser mayor. A su vez, teniendo en cuenta que la timidilato sintasa cumple una función que podría considerarse universal (participa en la biosíntesis de nucleótidos), tiene sentido que sea una enzima que se encuentre en organismos tan diversos que abarque desde bacterias hasta eucariontes.

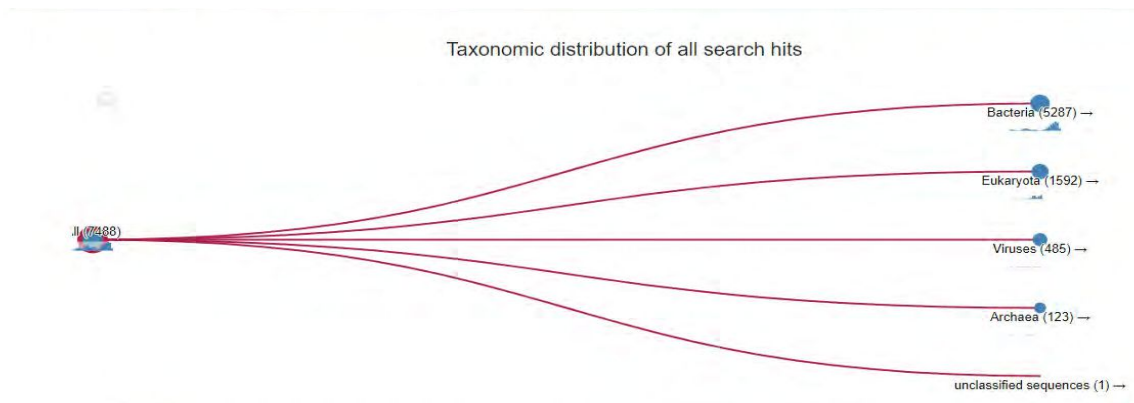


Figura 1. Árbol filogenético que se obtiene tras realizar la búsqueda secuencial utilizando HMMER.

A partir de los resultados obtenidos y teniendo en cuenta las distintas metodologías utilizadas podemos observar cómo se complementan las tres herramientas. Por un lado el BLAST nos provee de información correspondiente a los organismos más cercanos evolutivamente mientras que las herramientas como PsiBlast o HMMER permiten obtener homólogos más distantes con porcentajes de identidad menores pero con buenos E-valores.

Predicción de estructura secundaria, segmentos transmembrana, regiones desordenadas y dominios

Dominios

A partir de la base de datos de Pfam se encontró que toda la secuencia de la proteína corresponde a un único dominio correspondiente al dominio de la enzima timidilato sintasa y usando CDART encontramos que ese dominio sólo se encuentra en un tipo de arquitectura (compuesta sólo por el dominio Timidilato Sintasa) y en una única especie (*Candidatus Poseidoniales archaeon*). Predicción de estructura secundaria utilizando la herramienta QUick2D (<https://toolkit.tuebingen.mpg.de/tools/quick2d>).

Como se muestra en la *Figura 2* podemos observar que la estructura del dominio se basa en una combinación de alfa hélices y hojas beta y no contiene regiones coiled-coils ni transmembrana. A su vez, si bien esta herramienta muestra posibles coincidencias de regiones desordenadas cercanas al N-terminal, más adelante en el informe se analizará con otros programas las regiones desordenadas al igual que los segmentos transmembrana.

Segmentos transmembrana

Para el análisis de los segmentos transmembrana se utilizaron dos herramientas distintas: Phobius y HMMTOP2. Como se observa en la *Figura 3A*, no hay predicción de segmentos transmembrana y gráficamente se puede observar que la probabilidad de que el segmento comprendido entre los aminoácidos 150-200 no supera el umbral correspondiente al citoplasmático ni al no-citoplasmático (extracelular). Los resultados de la *Figura 3B* son resultados complementarios para corroborar que la Timidilato Sintasa de *Candidatus Poseidoniales* no posee regiones transmembrana. Además, los resultados obtenidos se condicen con lo encontrado en el servidor Toolkit para predicción de estructura secundaria. Con estos resultados podemos decir que la Timidilato Sintasa es una proteína citoplasmática.

Protein ID: tr|A0A6V8D8A1|A0A6V8D8A1_9EURY Thymidylate syntha

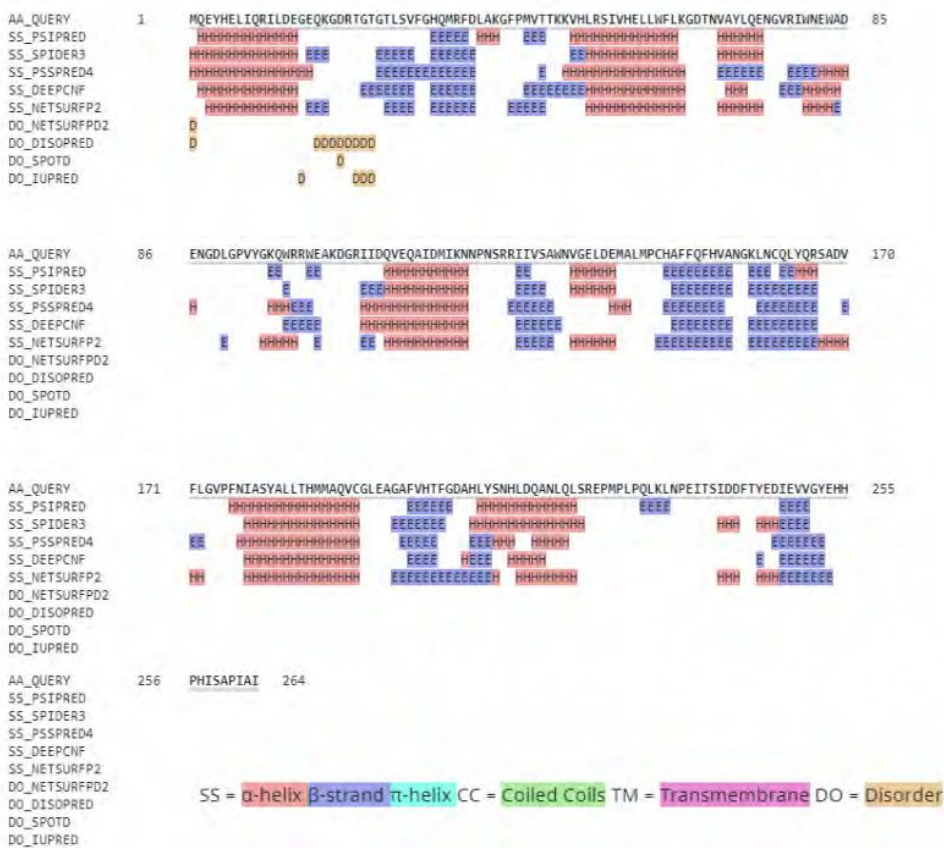


Figura 2. Resultados que se obtienen luego de correr el Quick2D. Las porciones rosas corresponden a las regiones alfa hélice, las azules a Hojas beta y las naranjas a las regiones desordenadas.

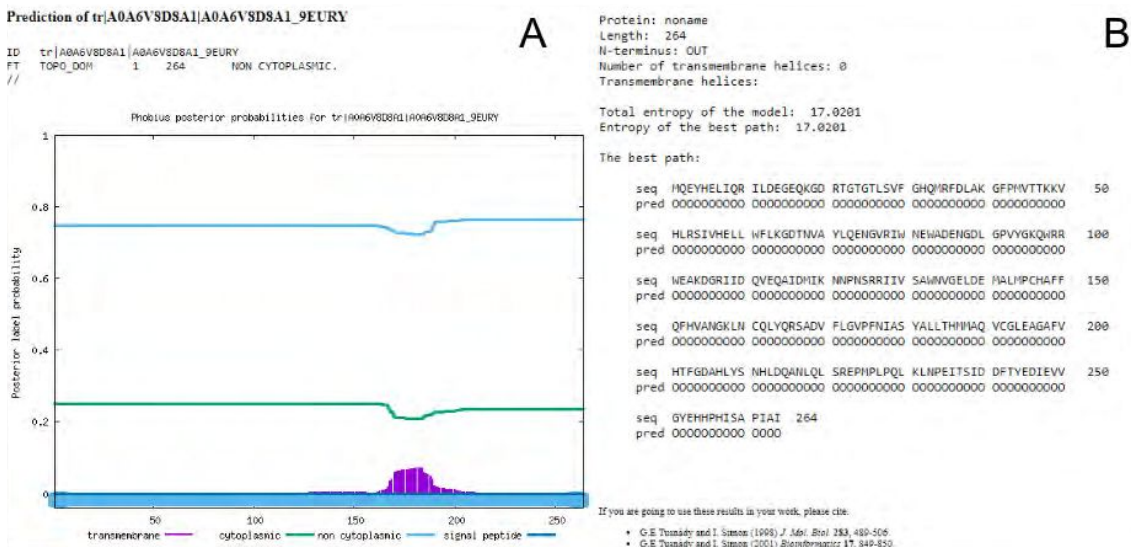


Figura 3. Resultados de la evaluación de segmentos transmembrana. A: resultado obtenido utilizando la herramienta Phobius. B: resultados obtenidos a través de HMMTOP2

Regiones desordenadas

A partir de la herramienta Dynamine se analizó la secuencia en búsqueda de regiones desordenadas y los resultados se muestran en la Figura 4. Se puede observar que la mayoría de las predicciones para cada aminoácido se encuentran por encima del umbral de flexibilidad, en otras palabras, gran parte de la

secuencia se considera rígida. Teniendo en cuenta que Dynamine asocia la flexibilidad con la presencia de regiones desordenadas, la presencia de rigidez entonces se corresponde con estructuras ordenadas. En conclusión, el dominio Timidilato sintasa es un dominio mayormente ordenado, con presencia de “desorden” en sus extremos N-terminal y C-terminal. Esto último tiene sentido si tenemos en cuenta que la mayoría de las proteínas tienen regiones más flexibles en sus extremos.

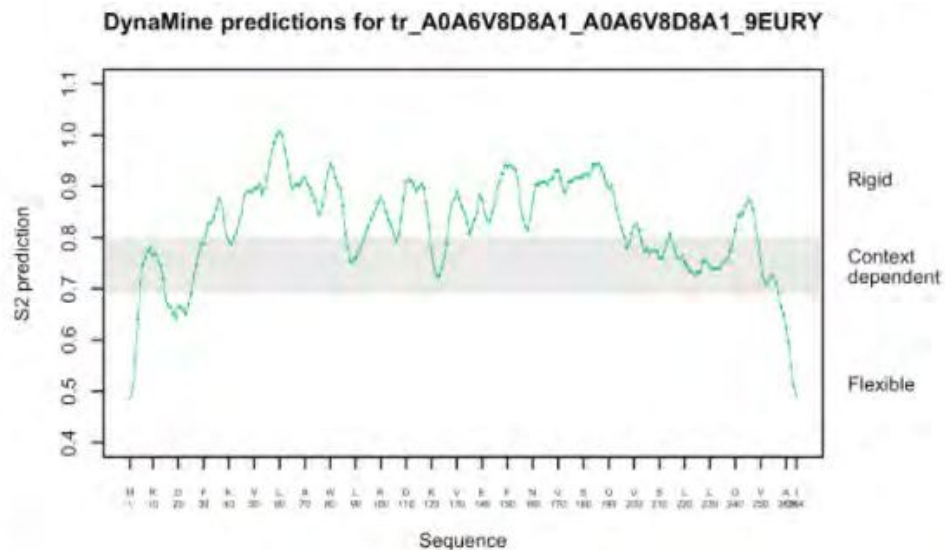


Figura 4. Resultados obtenidos utilizando la herramienta Dynamine
<http://dynamine.ibsquare.be/tag/disorder/>

Clasificación estructural de la proteína

Con el objetivo de seguir con la caracterización de la enzima Timidilato Sintasa de *Candidatus Poseidoniales archaeon* se evaluó la presencia de dominios estructurales pero a partir de herramientas que se basan en la aproximación de la estructura por homología de secuencia. Para ello se utilizó la base de datos de dominios estructurales CATH de la cual se obtuvieron los siguientes resultados: 231 dominios alineados y 32 matching Fun Fams.

Dentro de los 231 dominios, la clasificación CATH encontrada es la siguiente:

- C (clase): 3. Alpha- Beta
- A (arquitectura): 3.30. 2 Layer-Sandwich
- T (topología): 3.30.572. Thymidylate Synthase; Chain A
- H (Superfamilia de homólogos): 3.30.572.10. Thymidylate synthase/dCMP hydroxymethylase domain

Estos resultados concuerdan con los encontrados utilizando herramientas de predicción de estructura secundaria, pero a su vez da información de cómo se relacionan esas estructuras secundarias entre ellas y, de tenerlos, con otros dominios. En este caso, la información de proteínas con estructura conocida refuerza la idea de que la Timidilato Sintasa de *Candidatus Poseidoniales archaeon*, al igual que muchas de las Timidilato Sintasas, están compuestas por un único dominio.

Obtención de un modelo molecular para la proteína

Elección del Template

Se utilizan varias herramientas para seleccionar el mejor template posible, con un alto %Identidad, bajo E-value, buena cobertura, con estructura conocida y en lo posible una resolución menor a 2 amgstroms. El resultado que cumple con los requisitos mencionados se encontró utilizando HHPred y posee las siguientes características:

- Nombre: Thymidylate synthase (Escherichia coli (strain K12))
- Código PDB: 6NNR
- E-value: 3.4e-57
- %Identidad: 71%
- Similitud: 1.241
- Score: 393.2
- Posiciones alineadas: 264
- Resolución: 1.05 Å

Alineamiento: se lleva a cabo utilizando Clustal y se convirtió de manera manual al formato PIR.

Determinación de regiones conservadas y variables

Para generar el modelo tridimensional de la proteína se utilizó el programa Modeller que construye el backbone de carbonos alfa, los loops y cadenas laterales que luego podrán ser optimizadas. Para poder obtener un modelo tridimensional de la proteína de interés, al programa se le debe ingresar un archivo de comandos que indique el alineamiento y el template que debe utilizar para el modelado. El resultado son una serie de modelos con distintos valores de función objetivo. En este trabajo se eligió el modelo trA0A6V8D8A1.B99990003.pdb, que pondera entre el valor más bajo de la función objetivo y el DOPEscore. Cabe aclarar que el Modeller, a diferencia de otros programas, no utiliza una función objetivo con base energética, sino que a partir de la información del template genera una matriz con información estructural y a partir de la misma obtiene los “restrains”. Los “restrains” son parámetros estructurales que poseen un rango en el que pueden moverse. El mejor modelo es el que infringe la menor cantidad de restrains y por eso se trata de elegir el menor valor de función objetivo.

Evaluación del modelo en función del criterio energético utilizando ProsaII

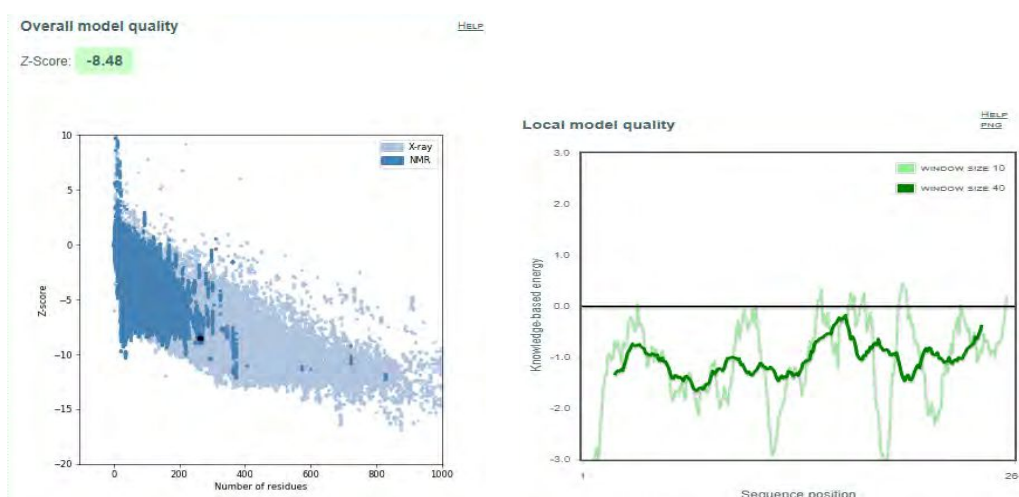


Figura 5. Izquierda: Resultados de Z-score en función del número de residuos de las proteínas con estructura conocida que se obtiene en ProsaII. El punto negro indica la ubicación del modelo. Derecha: Energía de cada posición dentro de la secuencia de interés.

La evaluación del modelo se puede llevar a cabo siguiendo criterios estructurales y/o energéticos. En este caso sólo se evaluó a partir de criterios energéticos y para ello se utilizó la herramienta ProsaII. En la Figura 5 se observa la evaluación del modelo a nivel global. El gráfico refleja la relación entre los valores de

Z-score, que se obtienen para proteínas con estructura conocida, en función de la longitud de cada proteína. Como el modelo se encuentra dentro de la distribución de posibles estructuras, podemos decir que es un buen modelo. Por otro lado, se muestra la distribución energética de cada posición de la secuencia del modelo (*Figura 5, derecha*), dando información de las regiones que pueden ser optimizadas, teniendo en cuenta que las posiciones de mayor energía corresponden generalmente a los loops. Como en este caso, las posiciones de mayor energía ni siquiera llegan a valores positivos y cómo esas regiones no se corresponden en loops, no se realiza una optimización de estos.

Visualización del modelo y comparación con el template en Pymol

En la *Figura 6* se observa el modelo obtenido alineado con el template 6NNR (Timidilato Sintasa de *E.coli*), que a simple vista, no demuestra tener grandes diferencias uno de otro y en la consola de Pymol se puede leer que el RMSD del alineamiento es de 0.18. Teniendo en cuenta que el error experimental asociado a la difracción de rayos X y al alineamiento de una misma proteína es de 0.5, podemos decir que el modelo obtenido a través de Modeller es muy similar al template si sólo tenemos en cuenta la información estructural.

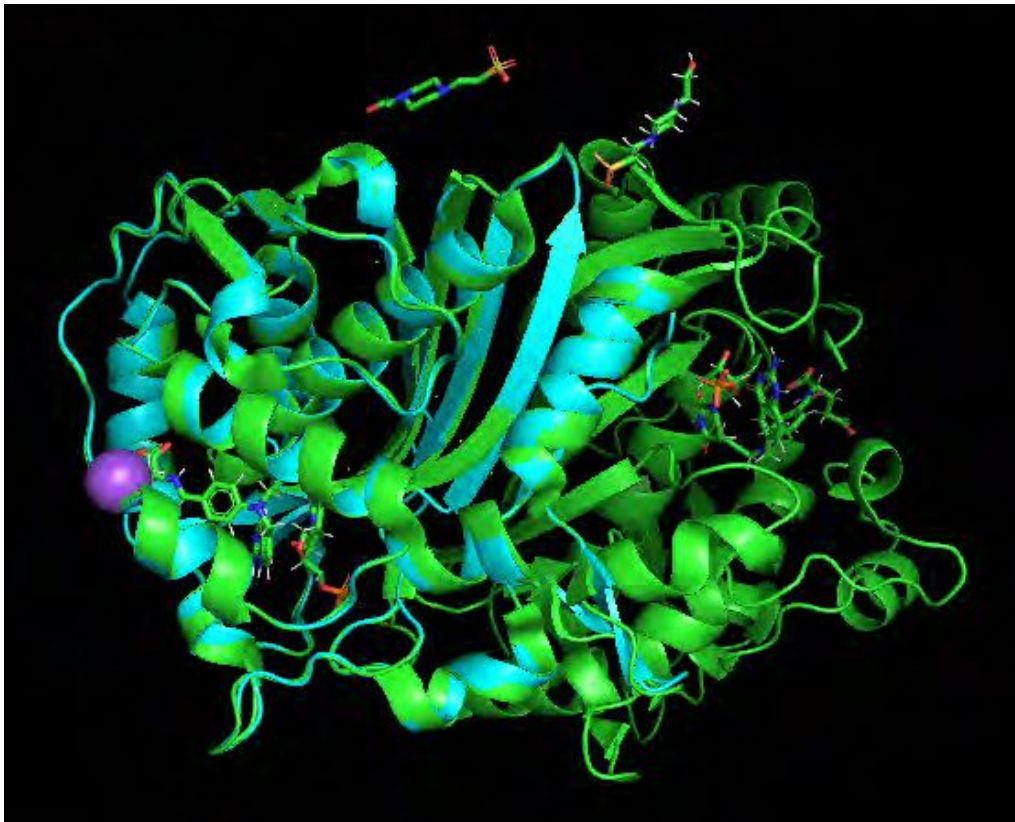


Figura 6. Superposición 3D del modelo (cian) y el template 6NNR (verde) que se obtiene a partir de Pymol.

Selección y alineamiento múltiple de secuencias

Con el objetivo de analizar las regiones secuenciales conservadas de la enzima Timidilato Sintasa y posterior análisis filogenético, se seleccionaron al azar 15 secuencias del BLAST con diferentes porcentajes de identidad que sean lo más heterogéneos posibles (Ver *Tabla 1* en el anexo). Se buscaron las secuencias de cada organismo en la base de datos de NCBI y a partir de las mismas en formato FASTA se procesaron utilizando la herramienta CLUSTAL <https://www.ebi.ac.uk/Tools/msa/clustalo/>.

A partir de las secuencias se observó que la longitud difiere entre los diferentes organismos. Esto puede deberse a que como la Timidilato Sintasa es un dominio completo, una secuencia más larga puede corresponder a una proteína con arquitectura compuesta por varios dominios donde uno de ellos es el

correspondiente a la Timidilato Sintasa. Dicho esto, y si bien el alineamiento no es de los mejores ya que presenta muchos gaps, donde comienza el alineamiento de la secuencia de interés se observan regiones con un grado de conservación considerable, teniendo en cuenta que se tomaron secuencias desde un 95% de identidad hasta un 14% de identidad. (Ver Figura 1 del anexo)

Si consideramos que la conservación secuencial conlleva a la conservación estructural y funcional, tiene sentido que la mayor parte del dominio se encuentre conservado, ya que de esa forma se asegura que la función se mantiene a lo largo del tiempo. Además, se observan algunas regiones que se encuentran más conservadas que otras (entre los aminoácidos 174-255) que con un futuro análisis estructural y funcional se podría llegar a dilucidar el rol que cumple esa región para que la secuencia se conserve más que en otras regiones.

Obtención de un árbol filogenético

Con el objetivo de obtener un árbol filogenético se utilizó la herramienta IQtree, que en primera instancia realiza una evaluación de modelos para encontrar el modelo de evolución que mejor ajuste al alineamiento múltiple presentado anteriormente. El modelo de evolución encontrado es LG y es el que se utiliza para estimar la topología del árbol filogenético. Además, se realizó un bootstrapping ultrarápido con 1000 iteraciones (alineamientos realizados al azar). Se obtuvo un árbol filogenético en formato newick que se usó como input en el programa iTOL que nos permite una mejor visualización.

Como se puede observar en la *Figura 7*, a medida que disminuye la similitud secuencial, es decir el porcentaje de identidad baja, las diferencias evolutivas aumentan. Por otro lado, se observa que a medida que disminuye la similitud secuencial, los valores de bootstrapping disminuyen, demostrando que la información que posee el alineamiento no es suficiente para soportar las inferencias filogenéticas en torno a la enzima Timidilato Sintasa cuando la similitud secuencial disminuye. Como era de esperarse, la secuencia Query corresponde a la misma Timidilato Sintasa de *Candidatus Poseidoniales archaeon*, y además la mayor cercanía evolutiva se da con proteínas con alto porcentaje de identidad. Salvo *Lasallia pustulata* que es un hongo y *Delftia phage RG-2014* y *deoxyuridylate hydroxymethyltransferase [Acinetobacter phage SH-Ab 15599]* que son fagos, el resto de las secuencias analizadas corresponden a proteínas de bacterias. Esta información se ve reflejada tanto en la similitud secuencial como en la divergencia evolutiva. La Timidilato Sintasa de *Lasallia pustulata* se encuentra más lejana evolutivamente que el resto de las enzimas de sus respectivos organismos.

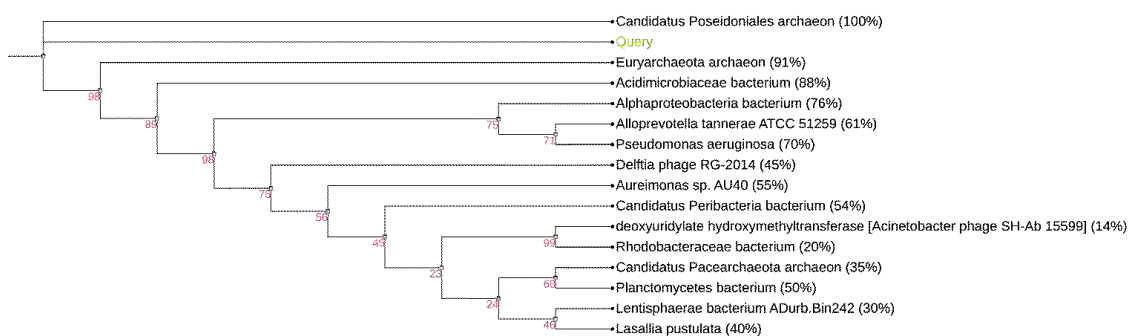


Figura 7. Árbol filogenético obtenido a partir de iTOL. Los números rosas corresponden a los valores de bootstrapping y los números entre paréntesis hacen referencia al porcentaje de identidad que comparten la secuencia de Timidilato Sintasa de ese organismo alineada con la secuencia Query.

Si bien no se pueden realizar inferencias filogenéticas a nivel de organismo ya que sólo se está analizando una única secuencia correspondiente a una única proteína, encontramos una mayor distancia evolutiva entre Timidilato Sintasas de organismos que presentan mayores diferencias (hongos-arqueas).

Análisis funcional de la Timidilato Sintasa

La estimación de la función de una proteína no es tarea sencilla y deben utilizarse la mayor cantidad de criterios para llegar al mejor resultado posible, que luego deberá ser confirmado de manera experimental. Desde el punto de vista evolutivo, resulta interesante evaluar las posiciones conservadas ya que estas posiblemente estén relacionadas con la función biológica de la enzima. Para ello se utilizó la herramienta Consurf que, mediante Evolutionary Trace, da información de las posiciones más conservadas y menos conservadas.

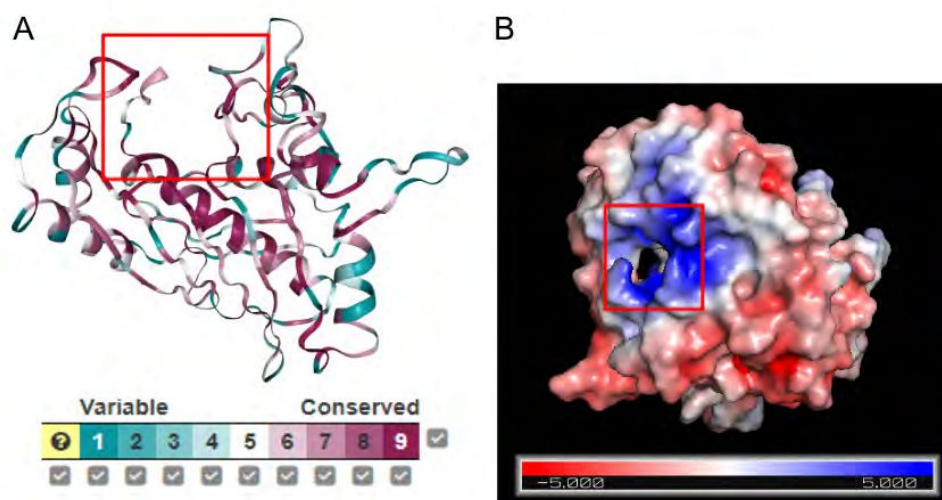


Figura 8. Representaciones espaciales del modelo de Timidilato Sintasa. **A.** Posiciones conservadas dentro del modelo. **B.** Carga superficial del modelo. Los recuadros rojos hacen referencia a la misma región y muestra un posible sitio de unión o entrada de sustratos.

Como se observa en la *Figura 8A*, las posiciones más conservadas se observan en la alfa hélice del centro de la cadena y en algunos loops que se encuentran encuadrados en rojo. Esa misma región en la *Figura 8B* muestra la presencia de un canal que puede llegar a ser importante en la funcionalidad de la proteína.

Al comparar la información del template utilizado en el modelado con el modelo obtenido a través de Modeller se puede llegar a una aproximación de cuál es el sitio activo de la proteína. Por un lado, se analiza la carga superficial del template utilizando Pymol (*Figura 9A*), donde se puede ver que los ligandos se encuentran inmersos en un canal conformado por aminoácidos de carga positiva. A su vez, la *Figura 9B* muestra la disposición estructural de la Timidilato Sintasa de *E.coli* (6NNR). Como puede observarse, la cavidad formada por aminoácidos de carga positiva (*9A*) direcciona hacia el sitio activo de la proteína, donde va a ingresar el (6R)-5,10-methylene-5,6,7,8-tetrahydrofolate y el dUMP (ambos con carga negativa), sustratos necesarios para que ocurra la reacción.

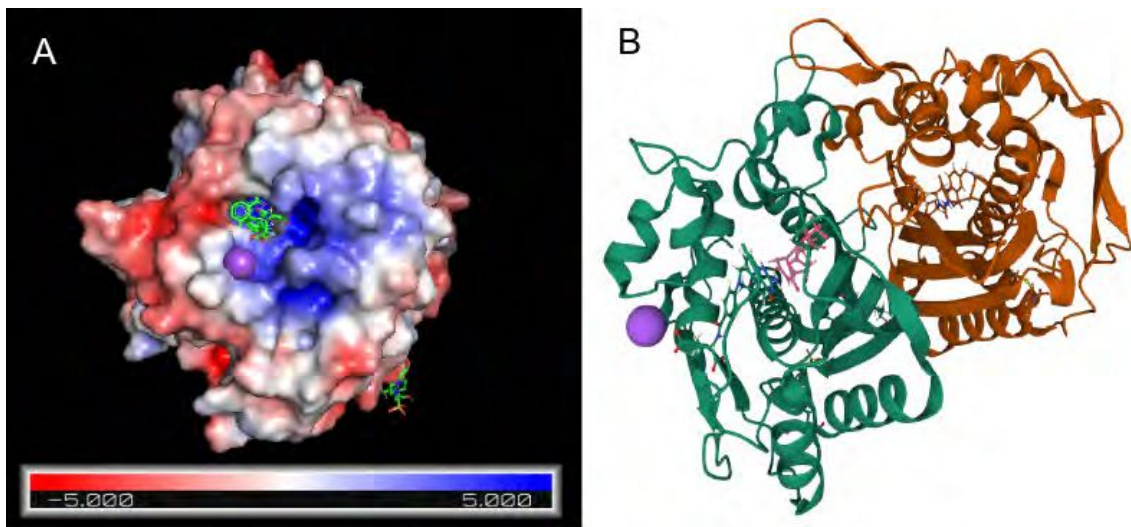


Figura 9. Estructura tridimensional de la Timidilato Sintasa de *E.coli*. **A:** carga superficial obtenida a partir de Pymol. **B:** Estructura 3D que se puede encontrar en la base de datos PDB. En rosa el dUMP y en azul (6R)-5,10-methylene-5,6,7,8-tetrahydrofolate, ambos sustratos de la reacción catalizada por la enzima.

El mismo procedimiento se realizó con el modelo obtenido en Modeller. Como se puede observar en la *Figura 10*, la carga superficial del canal varía levemente en el interior. En lugar de estar conformado en su totalidad por aminoácidos con carga positiva, se ve que hacia el interior aumenta la proporción de aminoácidos con carga negativa. Esto podría afectar a la función de la Timidilato Sintasa o simplemente puede estar reflejando que al ser de organismos distintos (el template proviene de *E.coli* mientras que nuestra proteína de un organismo del reino Archaea), su composición aminoacídica refleja una divergencia evolutiva correspondiente al hábitat diferencial en que normalmente se encuentran esos organismos.

Por otro lado, retomando la *Figura 8*, podemos confirmar que las posiciones más conservadas tienen una funcionalidad enzimática, ya sea que da lugar al canal de acceso al sitio activo o el sitio activo en sí.

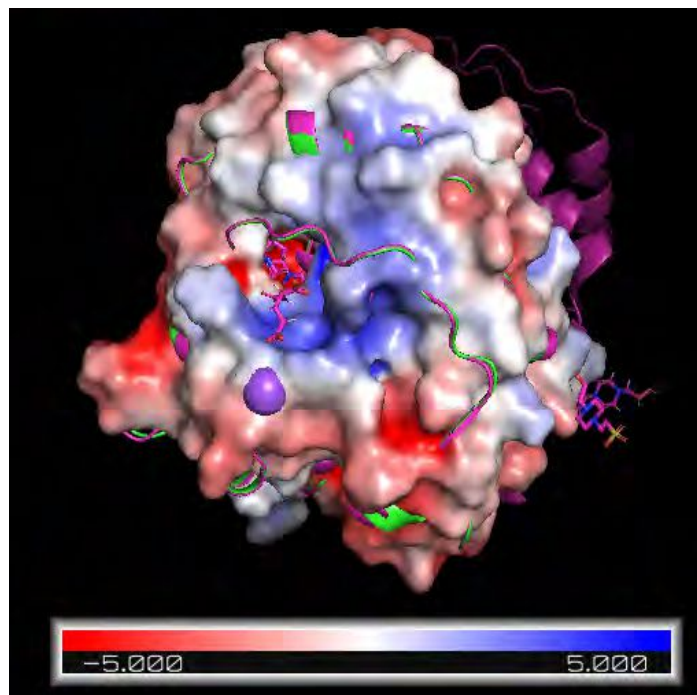


Figura 10. Carga superficial del modelo. El modelo(verde) está alineado con el template 6NNR (magenta)

Por último, en la base de datos de UniProt vemos que la proteína de interés tiene las siguientes anotaciones GO:

- Función Molecular: actividad metiltransferasa, actividad transferasa, actividad timidilato sintasa, actividad de transferencia de un grupo carbono
- Proceso Biológico: metilación, proceso de biosíntesis de nucleótidos, biosíntesis de dTMP.

CONCLUSIONES Y DISCUSIÓN

El uso de diversas herramientas bioinformáticas nos permitieron identificar y confirmar que la secuencia de interés corresponde a la enzima Timidilato Sintasa presente en una especie de archeas conocida como *Candidatus Poseidoniales archaeon*. A partir de BLAST encontramos secuencias homólogas evolutivamente cercanas mientras que PsiBlast y HMMER que utilizan otros algoritmos basados en la generación de perfiles y HMM, nos permitieron obtener secuencias homólogas más lejanas. A partir de estos resultados se puede buscar un gran número de propiedades y características estructurales que intervienen en la funcionalidad y es así como encontramos que la proteína de interés se encuentra en el citoplasma y está conformada por un único dominio. A su vez ese dominio posee una estructura que se basa en la combinación de alfa hélices y hojas beta donde las regiones desordenadas se encuentran en las regiones terminales de la proteína.

La información secuencial nos permitió obtener la relación evolutiva entre secuencias proteicas provenientes de distintos organismos y de esa manera pudimos realizar inferencias funcionales de los canales por los que ingresan los metabolitos necesarios para que ocurra la reacción. También encontramos que si bien las regiones funcionales de la Timidilato Sintasa se conservan a lo largo de varias especies, el cambio de aminoácidos de carga positiva por aminoácidos neutros o de carga negativa pueden correlacionarse con el ambiente en el que se encuentran y desarrollan los diferentes organismos.

Resulta interesante evaluar la diversidad y la flexibilidad que presentan las herramientas utilizadas a lo largo del trabajo. Tanto la hipótesis como la pregunta biológica que se desea responder determinan los programas o herramientas a utilizar y qué información se utilizará como input. En este trabajo nos planteamos como objetivos corroborar la información brindada por las diferentes bases de datos sobre la Timidilato Sintasa de *Candidatus Poseidoniales archaeon* y a su vez explorar el modelado tridimensional de dicha enzima y la importancia de las secuencias conservadas sobre la funcionalidad de la misma y para ello se utilizaron herramientas que tienen un enfoque evolutivo que nos permitieron inferir o hipotetizar sobre los objetivos planteados. Es importante tener en cuenta que la inferencia bioinformática no reemplaza la necesidad de corroborar las hipótesis con medidas experimentales.

Por último, en el caso particular de la Timidilato Sintasa de *Candidatus Poseidoniales* donde existe bastante información al respecto de su función y disposición estructural, con el uso de las herramientas presentadas se pudo obtener un modelo estructural de la proteína que hasta el momento no se tenía. Además, a través de información evolutiva y estructural se pudo reconocer el sitio activo de la proteína.

BIBLIOGRAFÍA

Rinke, C., Rubino, F., Messer, L.F. et al. A phylogenomic and ecological analysis of the globally abundant Marine Group II archaea (Ca. Poseidoniales ord. nov.). *ISME J* 13, 663–675 (2019). <https://doi.org/10.1038/s41396-018-0282-y>

Caracterización secuencial, estructural y evolutiva del receptor de Kisspeptina, GPR-54, en *Mus musculus* mediante herramientas bioinformáticas

Alejandro Raúl Schmidt

Cátedra de Bioinformática, Área de Biotecnología y Biología Molecular, Departamento de Ciencias Biológicas, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, La Plata, Argentina.

RESUMEN

La kisspeptina cumple un rol fundamental como regulador del eje reproductivo y se la ha asociado a un receptor acoplado a proteína G heptahelical de la familia de la rodopsina (GPR-54), ubicado en la membrana plasmática de las neuronas con expresión de la hormona liberadora de gonadotropina (GnRH), la cual es de vital importancia para la reproducción. Con el objetivo de caracterizar secuencial, estructural y evolutivamente a GPR-54, en *Mus musculus*, se utilizaron diversas herramientas bioinformáticas.

A partir del código Uniprot (Q91V45) se accedió a la secuencia FASTA y se realizó un BLAST en busca de homólogos dentro de Rodentia para su posterior alineamiento múltiple de secuencias que destacará la variabilidad y características clase específicas de los mismos. De esta forma, se obtuvieron 6 grupos de proteínas homólogas (receptor de Kisspeptina tipo 1, receptor de galanina tipo 1, 2 y 3, receptor de somatostatina tipo 4 y receptor opioide tipo kappa) con perfiles secuenciales particulares en el extremo N terminal y en el primero, quinto y sexto loop extracelular. En el mismo alineamiento se encontró una alta conservación de los residuos cuya mutación dispara la pérdida de función del receptor y que por análisis de Evolutionary Trace mostraron no ser de importancia funcional.

Estructuralmente se construyeron modelos por homología con el Modeller y ab initio vía trRosetta con rmsd de 1.3 y 0.8, respectivamente, que permitieron visualizar y confirmar las predicciones de estructuras secundaria generadas por el servidor QUICK2D, que establecían la presencia de alfa hélices en los 7 segmentos transmembrana.

Para finalizar se realizó un análisis filogenético utilizando como modelo evolutivo JTT + F, establecido por Modeltest, y PHYML, para la construcción de un árbol con las secuencias previamente estudiadas de los 6 grupos de proteínas homólogas. De este estudio, se destaca que el bootstrap del nodo de las GPR-54 con los receptores de galanina, que fueron las secuencias con mayor similitud, es de 65.

A modo de conclusión se logró caracterizar secuencial, estructural y evolutivamente a GPR-54, en una triada que se interrelaciona desde la perspectiva estructura-función y que es el reflejo del paso evolutivo.

PALABRAS CLAVE: kisspeptina; mamíferos; bioinformática

INTRODUCCIÓN

El eje neuroendocrino reproductivo es regulado de forma coordinada por los neuromoduladores GABA, Dopamina, Glutamato y Kisspeptina, entre otros, para garantizar la síntesis y liberación de GnRH (Lehman y col., 2010), la cual a su vez estimula la secreción de LH y FSH hipofisarios (Messenger y col., 2005; Navarro y col., 2005; Kauffman y col., 2007).

Entre los mencionados, kisspeptina (Kiss) es un péptido de 54 aminoácidos de expresión hipotalámica, el cual cumple un rol fundamental como regulador del eje reproductivo, pero que fue originalmente descrito en 1996 en líneas celulares de melanoma humano como un supresor de metástasis tumoral (Lee y col., 1996). Recién en 2001 Kiss se asoció con su receptor GPR-54 (Kotani y col., 2001; Muir y col., 2001; Ohtaki y col., 2001).

Dicho receptor fue descrito por primera vez en 1999 como un receptor huérfano acoplado a proteína G heptahelical de la familia de la rodopsina, clonado inicialmente de cerebro de rata (Lee y col., 1999) y posteriormente identificado en humano (Kotani y col., 2001; Clements y col., 2001; Muir y col., 2001; Ohtaki y col., 2001). GPR-54 consta de cinco exones que codifican una proteína de 396 aminoácidos en ratones, cuyo peso molecular es de 75 kDa (Civelli & Zhou, 2008).

Estructuralmente tiene un dominio N-terminal extracelular seguido por siete hélices transmembrana y el dominio citoplásmico C-terminal de aproximadamente 70 residuos (Seminara y col., 2003; Pasquier y col., 2014), el cual se une a las subunidades catalíticas y reguladoras de la fosfatasa 2A y forma complejos con las proteínas asociadas involucradas en la señalización del receptor (Bianco y col., 2013; Pasquier y col., 2014).

Su localización, caracterizada mediante hibridación in situ y RT-PCR cuantitativa, detectó altos niveles de expresión de GPR54 en cerebro, particularmente en el hipotálamo. A su vez, mediante hibridación in situ de doble marca y por inmunofluorescencia confocal GnRH / GPR54 se le reconoció un rol fundamental para la regulación del eje reproductivo (Messenger y col., 2005; Clarkson y col., 2008), pudiendo ratificarlo con estudios en ratones GPR54 K.O, que evidenciaron un impedimento para llegar a la pubertad, con órganos reproductores inmaduros y bajos niveles de esteroides sexuales y hormonas gonadotróficas. Finalmente, tanto en humanos como roedores, está descrito que las mutaciones L102P, L148S, C223R y R331X alteran la funcionalidad de GPR-54 y conducen a infertilidad por hipogonadismo hipogonadotrófico (Ohtaki y col., 2001; Kauffman y col., 2007; Civelli & Zhou, 2008).

OBJETIVOS

- Determinar la máxima distancia evolutiva mediante homólogos del receptor GPR-54.
- Establecer las diferencias y la conservación secuencial presentes entre las macromoléculas homólogas dentro de los principales filos de Rodentia (*Myomorpha*, *Hystricomorpha* y *Sciuromorpha*), destacando las posiciones importantes que causan infertilidad.
- Caracterizar estructuralmente a partir de predictores la secuencia del receptor GPR-54.
- Modelar el receptor GPR-54 de *Mus musculus* por homología y ab initio
- Establecer la filogenia de GPR-54 y sus homólogos dentro de Rodentia.
- Determinar la función de la estructura por medio de predictores.

MÉTODOS Y RESULTADOS

El trabajo realizado inició a partir del código Uniprot (Q91V45) proporcionado por los docentes y en dicha base de datos desde la sección sequence se descargó el siguiente archivo FASTA con la secuencia aminoacídica de 396 residuos:

```
>sp|Q91V45|KISSR_MOUSE KiSS-1 receptor OS=Mus musculus OX=10090 GN=Kiss1r PE=1 SV=1
MATEATLAPNVTWWAPSNASGCPGCGVNASDDPGSAPRPLDAWLVPFFATLMLLGLVGNLSVIYVICRHKHMQTTNFYIA
NLAATDVTFLCCVPFTALLYPLPAWVLGDFMCKFVNYIQVSVQATCATLTAMSVDRWYVTVFPLRALHRRTPRLALAVLSLI
WVGSAAVSAPVLALHRLSPGPRTYCSEAFPSRALERAFALYNLLALYLLPLLATCACYGAMLRHLGRAAVPAPT DGALQGQLL
AQRAGAVRTKVSRLVAAVVLLFAACWGPIQLFLVLQALGPSGAWHPRSYYAVKIWAHCMSYSSALNPLLYAFLGSHFRQA
FCRVCPCCRQRQRRPHTSAHSDRAATHVPHSRAAHPVIRIRSEPEGNPVRVSPCAQSERTASL
```

A partir de la secuencia se realizó un Blast en el NCBI en búsqueda de homólogos y para ello se utilizó una base de datos non redundant que incluyó GenBank CDS translations + PDB + SwissProt + PIR + PRF, y los parámetros de corte E-Value 0.00005, coverage 70% y porcentaje de identidad al 30%, se recuperaron 4308 secuencias.

De estas el homólogo más cercano se encontró dentro del mismo género en la especie *Mus caroli*, en la cual kiSS-1 receptor isoform X1 presentó 98% de identidad y similitud, sin inserción de gaps. En el otro extremo de la distribución de homólogos con un 30% de identidad, 50% de similitud y 6% de gaps encontramos el receptor de galanina tipo 3 en la especie *Cyanistes caeruleus*, la cual es un ave de la familia Paridae.

Con la idea de encontrar homólogos más lejanos y ante la imposibilidad técnica de realizar un PSIBLAST con la base de datos non redundant (nr, NCBI) completa se decidió utilizar este método iterativo con la base de datos landmark (NCBI) que incluye a los organismos modelo. Los parámetros utilizados en este caso fueron E-Value 0.00005 y hasta 3 iteraciones, encontrando en este caso como homólogo más lejano a la rhodopsina 5 en su isoforma B en *Drosophila melanogaster* con 20% de identidad, 35% de similitud, 87% de cobertura y 3% de gaps.

Como se planteó el objetivo de establecer diferencias y conservación secuencial de GRP-54 dentro de los principales subórdenes de Rodentia, se decidió filtrar la salida que se obtuvo por Blast desde la base de datos Non redundant para *Myomorpha*, *Hystricomorpha* y *Sciuromorpha*, siguiendo también con la utilización de los parámetros de coverage 70% y porcentaje de identidad al 30%. En este estudio se recuperaron 126 secuencias, siendo el homólogo más lejano encontrado el receptor de galanina tipo 2 en la *Peromyscus leucopus*, de la familia Cricetidae con un porcentaje de identidad de 30.48%, similitud 49.29% y 6% de gaps.

Todas las secuencias obtenidas se alinearon usando T-Coffee (vía EMBOSS) y el alineamiento obtenido se visualizó con los programas Jalview y GenDoc.

A partir del alineamiento múltiple de las 126 secuencias recuperadas, se evidenció la presencia de 6 grupos de proteínas homólogas (receptor de Kisspeptina tipo 1, receptor de galanina tipo 1, 2 y 3, receptor de somatostatina tipo 4 y receptor opioide tipo kappa) las cuales mostraron un patrón en la secuencia particular para cada una de ellas.

Dado que dentro de estos 6 grupos no se observaron variaciones mutacionales mayores, se decidió seleccionar al azar 2 secuencias representativas de cada grupo y realizar un nuevo alineamiento que será representativo de estos.

En el extremo N terminal de los homólogos alineados se observan grandes diferencias en el largo de secuencia característico de cada tipo de proteínas, esto es debido a que el algoritmo inserta gaps para mejorar el alineamiento. Así mismo no se observa conservación aminoacídica en este segmento secuencial (figura 1).

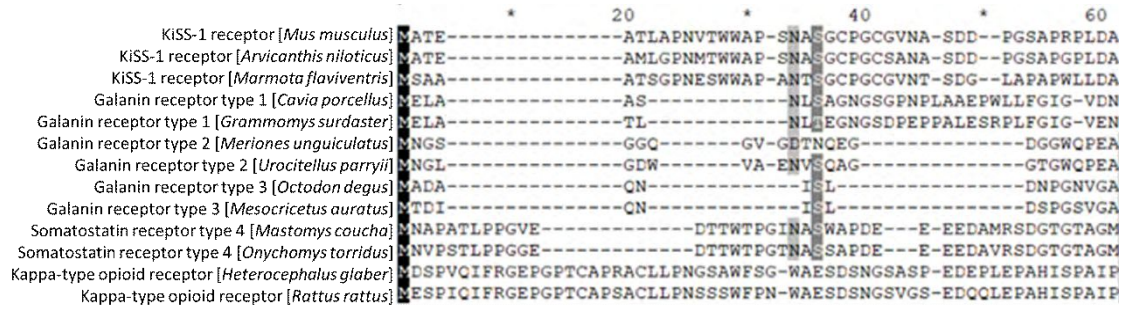


Figura 1. Alineamiento múltiple de secuencia del extremo N terminal de Kiss-1R y sus homólogos, visualizado por GenDoc.

Ante el conocimiento que GPR-54 es una proteína de 7 pasos transmembrana y se contaba con la posición de cada uno de ellos desde la información brindada por Uniprot, se evaluó la conservación de residuos tanto a nivel intersegmentos como de los segmentos transmembrana propiamente dichos.

En el primero de los casos se observan inserciones reflejadas por los gaps que modifican el largo de los loops (figura 2). En el receptor de galanina tipo 2 y 3 hay inserciones en el primer loop extracelular, mientras que en el receptor de Kisspeptina 1 las inserciones se observan en los últimos 2 loops, citoplasmático y extracelular.

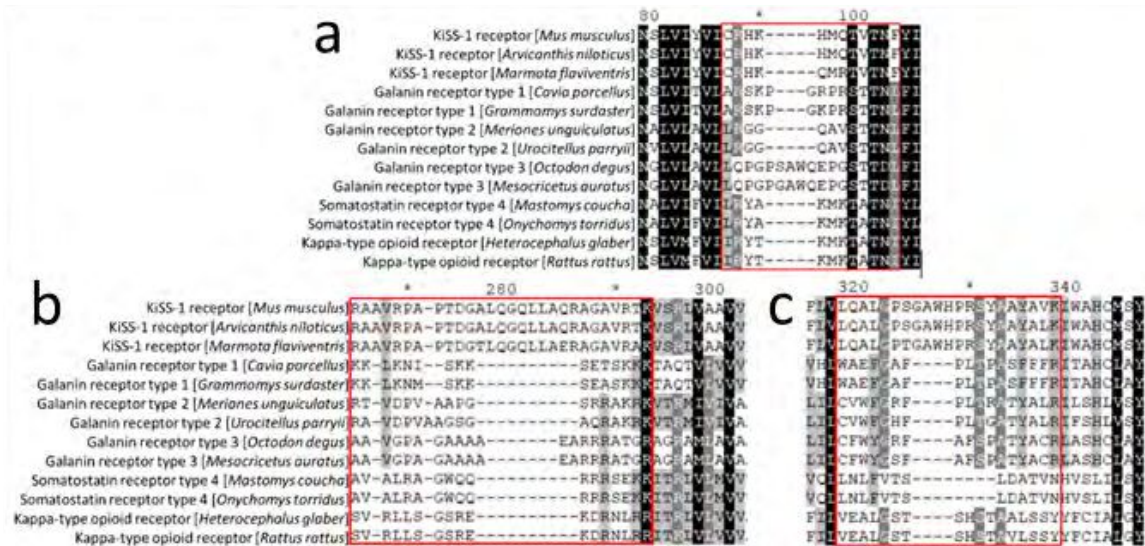


Figura 2. Alineamiento múltiple de secuencia de intersegmentos transmembrana de Kiss-1R y sus homólogos, visualizado por GenDoc. Cuadro rojo, destaca los residuos que componen el intersegmento en las diferentes secuencias. a. Primer loop extracelular. b. quinto loop citoplasmático. c. sexto loop extracelular.

Dentro de los 7 segmentos transmembrana informados en Uniprot, encontramos en todos los casos la presencia de aminoácidos altamente conservados, mayormente de propiedad fisicoquímica hidrofóbica. A modo de ejemplo en la figura 3 se muestra el segmento que va del aminoácido 79 al 101 de la secuencia query en donde se destaca que aproximadamente 2/3 de los residuos son neutros no polares o hidrófobos (alanina, valina, leucina, isoleucina, metionina, prolina y fenilalanina) y 1/3 neutros polares o hidrófilos (glutamina, asparagina, tirosina y cisteína).

Cabe destacar que dentro de una posición parcialmente conservada, los cambios mutacionales observados serían específicos de una rama evolutiva, como ser la posición 103 (flecha de la figura 3) del alineamiento en donde se observa que todos los receptores de kiss-1 poseen fenilalanina, mientras que los receptores de galanina leucina y finalmente los receptores de somatostatina y opiodes tipo kappa isoleucina.

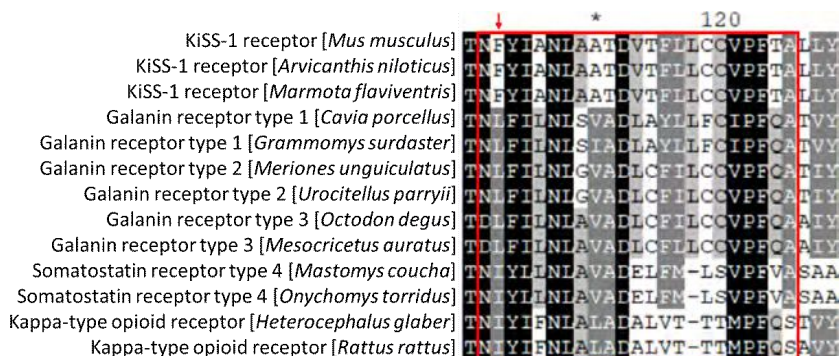


Figura 3. Alineamiento múltiple de la secuencia de Kiss-1R y sus homólogos, visualizado por GenDoc. Cuadro rojo, destaca los residuos que componen el segundo segmento transmembrana en las diferentes secuencias. Flecha roja, indica la posición 103.

Esta característica secuencial de gran cantidad de posiciones completamente conservadas, una mayor proporción de aminoácidos de carácter hidrofóbico y mutaciones sitio específicas que se comparten entre las proteínas de un grupo, se repite en el resto de los segmentos transmembrana.

Otra particularidad secuencial a destacar es que el último 10% del alineamiento correspondiente al extremo C-terminal, no presenta conservación secuencial salvo las posiciones prolina 437 y 443 que se encuentran parcialmente conservadas (figura 4).

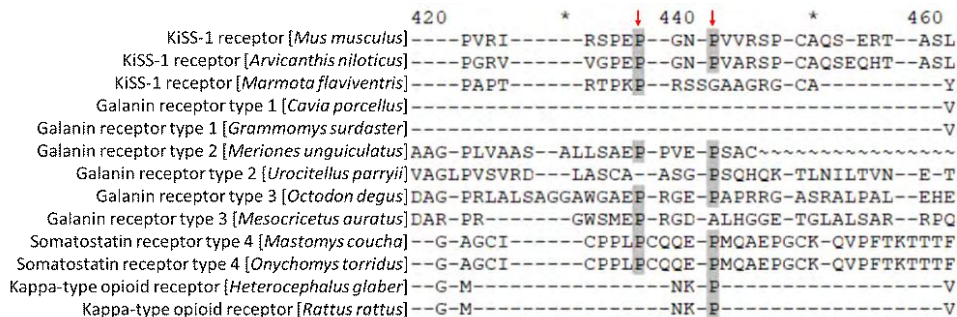


Figura 4. Alineamiento múltiple de secuencia de Kiss-1R y sus homólogos, visualizado por GenDoc. Se observa el extremo C-terminal del alineamiento. Flechas rojas indican las posiciones de prolina 437 y 443.

Con respecto a las posiciones y residuos fundamentales para la función de GPR-54, se puede observar que las posiciones L102, L148 y C223 del loop y la posición R331 citoplasmática se encuentran completamente conservadas en el grupo de receptores Kiss-1, y parcialmente, con sustituciones por aminoácidos mayormente de la misma característica fisicoquímica, en las proteínas homólogas (figura 5).

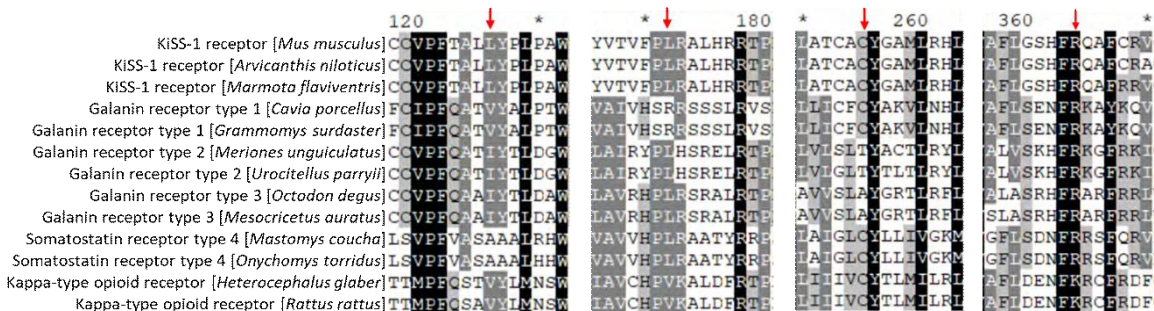


Figura 5. Alineamiento múltiple de secuencia de Kiss-1R y sus homólogos, visualizado por GenDoc. Se observan las posiciones L102, L148 y C223 del loop y la posición R331 el extremo C-terminal indicadas con flechas rojas.

Para completar la caracterización secuencia de GPR-54 (Q91V45) se procedió a su análisis mediante predictores de la estructura secundaria y asociados. Como primera medida, se corrió a partir de la secuencia de GPR-54 el servidor online Quick2D, el cual integra diferentes predictores de estructura secundaria de tercera generación (basados en redes neuronales), sumando a su salida información de la identificación de segmentos transmembrana y aminoácidos que participan de una región desordenada (figura 6).

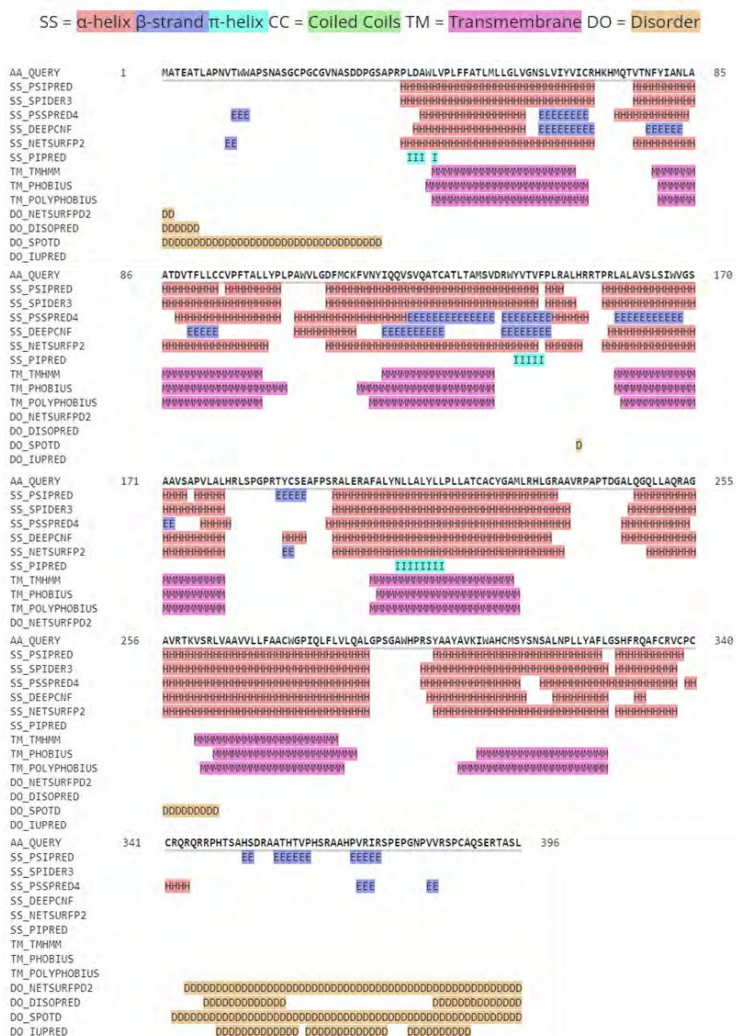


Figura 6. Salida del Quick2D, donde se observa la secuencia aminoacídica de GPR-54 y los elementos de la estructura secundaria, segmentos transmembrana y residuos que participan del desorden predichos.

A partir de la figura 5, se puede observar que la gran mayoría del resto de la proteína se compone de bloques de Hélices α ordenados dentro de 7 segmentos que coinciden con segmentos transmembrana. En ninguno de los 7 segmentos las Hélices α forman parte de coiled coils.

En cuanto a la estructura secundaria hoja β , a pesar de encontrarse presente en aminoácidos transmembrana de algunos predictores, esto fue descartado por consenso, pero se hipotetiza la presencia de algunos aminoácidos que puedan adoptar esta estructura en el extremo C terminal de la proteína.

En cuanto a la presencia de regiones desordenadas en los extremos amino y carboxilo terminal, se observó consenso de predicción de desorden mediante diferentes programas en el extremo carboxilo, pero no así en el amino terminal. Para profundizar en la caracterización de las regiones desordenadas se utilizó la base de datos MobiDB. Los resultados obtenidos permiten corroborar la presencia de desorden en el extremo carboxilo terminal, no así en el extremo amino terminal, a pesar de presentar aminoácidos tendientes al desorden (línea MobiDB-lite). En el apartado estructura secundaria utiliza la plataforma FeSS, la cual destaca la presencia de un alto contenido de Hélices (línea Helix FaSS) que son coincidentes con la existencia del dominio estructural único establecido por homología (línea Domains consensus). Por último, al integrar MobiDB información de la base de datos de Uniprot, brinda en su salida la presencia de los 7 segmentos transmembrana predichos anteriormente por el QUICK2D (línea Transmembrana), ubicados entre los aminoácidos 44-66, 79-101, 117-138, 158-180, 204-224, 261-283 y 306-330 (figura 7).

Para caracterizar cuáles eran los dominios presentes en la proteína se recurrió a la base de datos Conserved Domain Database (CDD), redireccionada desde el BLAST de NCBI (figura 8).

Los resultados indican la presencia de un único dominio correspondiente al receptor de péptido derivado de KiSS-1, miembro de la familia de la clase A de receptores acoplados a proteína G de siete pasos transmembranas (código de acceso CDD: cd15095), perteneciente a la superfamilia 7tm_GPCRs de receptores acoplados a proteína G de siete pasos transmembranas (código de acceso CDD: cl28897).

Cabe destacar que la superfamilia es compartida entre todas las proteínas homólogas caracterizadas por Blast, entre las que se incluyen a los 6 grupos de receptores Kisspeptina tipo 1, galinina tipo 1, 2 y 3, somatostatina tipo 4 y opioide tipo kappa. La carencia de coiled coils fue corroborada por el software NCOILS, el cual utiliza un método de ventana para la predicción, utilizando ventanas de 14, 21 y 28 residuos. En ningún rango se hizo presente la estructura secundaria coiled coils.

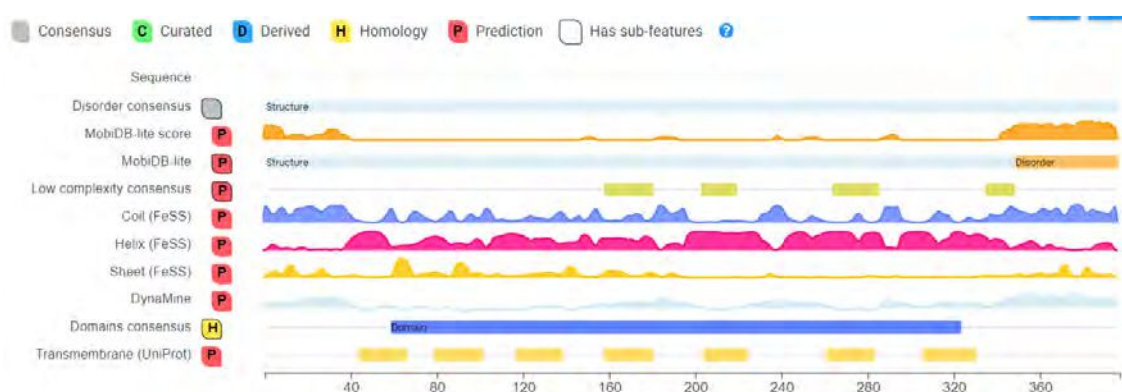


Figura 7. Salida del MobiDB para el código Uniprot Q91V45. La imagen integra secuencia consenso de desorden, estructuras secundarias predichas por FeSS, dominio consenso y elementos transmembrana.



Figura 8. Salida del Conserved Domain Database. Se observa un solo dominio perteneciente a la familia 7tmA_KISS1R de la clase A de receptores acoplados a proteína G de siete pasos transmembranas y las superfamilias 7tm_GPCRs, PHA03087 y 7tm_1.

En el caso del perfil de segmentos transmembranas, éste fue corroborado e ilustrado con el predictor TMHMM, basado en la presencia de Hélices α consecutivas (figura 9). Cabe destacar que el programa ha predicho correctamente la región outside-inside hasta el sexto segmento transmembrana, pero ha fallado en el reconocimiento del séptimo segmento debido a la baja probabilidad y en la secuencia carboxilo terminal intracitoplasmática al caracterizarla como outside en el global, al verse arrastrada por la probabilidad del último segmento.

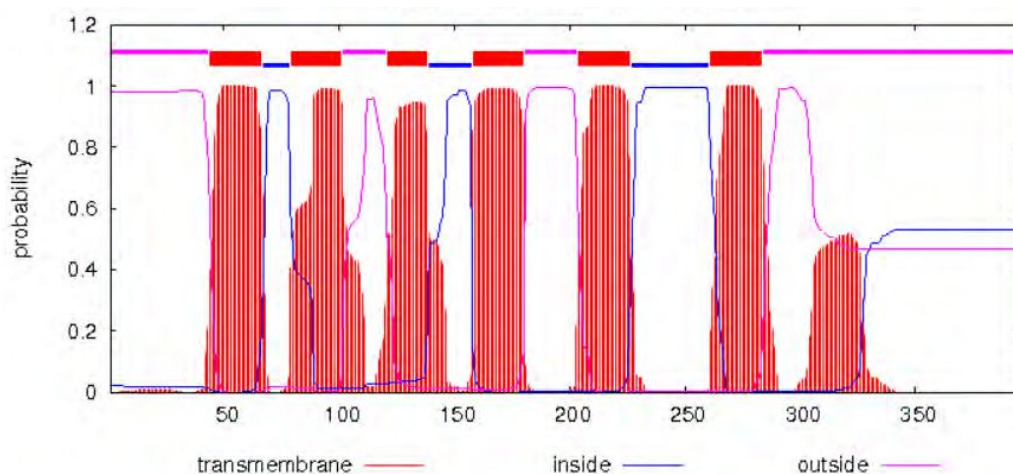


Figura 9. Salida del TMHMM 2.0. Se observan basados en probabilidad, los 7 segmentos transmembrana y los elementos secuenciales inside-outside.

Por último, en el análisis secuencial, se verificó la incongruencia de desorden en el extremo amino terminal por medio de uno de los predictores de desorden más sólidos para su estimación como lo es el Iupred2A. Aquí se encontró que si bien existe tendencia al desorden en el extremo amino no alcanza a llegar al cutoff que considera el programa (figura 10).

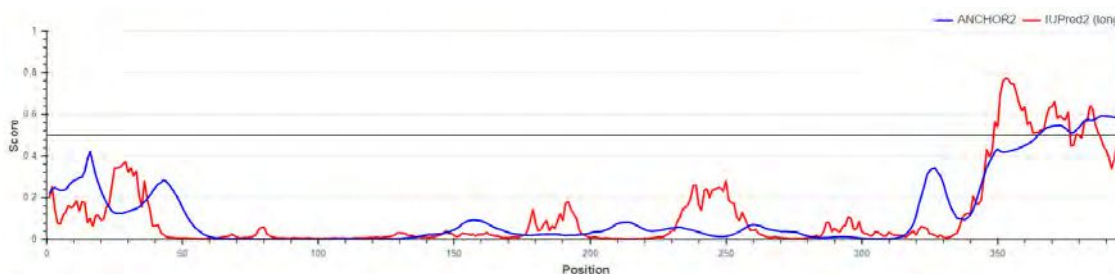


Figura 10. Salida del Iupred2A. El extremo carboxilo terminal de la secuencia de GPR-54 de *Mus musculus* presenta aminoácidos que superan el cutoff de desorden. La línea roja corresponde al desorden. Línea azul corresponde a las regiones con desorden que tienen capacidad de interacción con ligandos (figura 10).

Una vez analizada la secuencia, y en la búsqueda de profundizar el conocimiento estructural de GPR-54, se realizó un modelado por homología. Para el mismo, como primera medida, se establecieron los templates a partir de blast con la secuencia de Kiss-1R de *Mus musculus* y la base de datos de la PDB, dada la necesidad

de que las secuencias homólogas tengan el mejor porcentaje de identidad posible, una buena cobertura y estructura conocida.

De dicha búsqueda se obtuvieron dos posibles candidatos para el modelado, el receptor de orexina-1 (6TO7) y el receptor delta opioide 7TM (4N6H), ambos de *Homo sapiens* (tabla 1). El alineamiento de GPR-54 y los candidatos elegidos para el modelado se realizó con el programa T-Coffee.

Tabla 1. Homólogos de GPR-54 de *Mus musculus* candidatos para el modelado, con los datos de relevancia para el modelado por homología.

Código	Identidad (%)	Similitud (%)	Cobertura (%)	Resolución (Å)
6TO7	32.52	49	69	2.29
4N6H	31.96	50	72	1.8

El programa Modeller, se ejecutó localmente, para obtener 25 modelos con cada homólogo, los que fueron evaluados según los valores de Molpdf y DOPE.

Los mejores modelos para 6TO7 (el 13) y para 4NH6 (el 4) fueron evaluados mediante ProSA-web. Ambos mostraron no ser de calidad suficiente, con Z-scores que se escapaban a la distribución normal, además de contener regiones de alta energía. Cabe destacar que ninguno de los templates permitió un correcto modelado de las regiones amino y carboxilo terminal. Así mismo en pos de mejorar el modelo, se editó el alineamiento del mejor template, el 4N6H, y se eliminaron aminoácidos no alineados de los extremos amino y carboxilo terminal. Con esta metodología se logró una leve mejoría del modelo evidenciado por los análisis mediante ProSA, obteniendo un acercamiento del Z-score a la nube de distribución y disminuyendo la energía (figura 11).

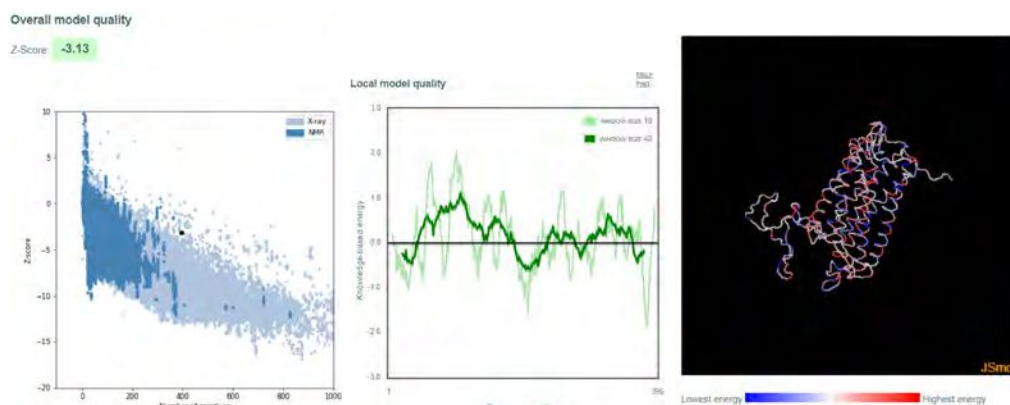


Figura 11. Salida de ProSA para el modelo de GPR-54 de *Mus musculus* desde 4NH6 editado.

Las estructuras del modelo PDB y el modelado por homología de GPR-54 fueron alineadas utilizando el software PyMOL para su visualización (figura 12) y a través del servidor DALI para evaluar la disimilitud estructural mediante el valor de rmsd que fue de 1.3 Å. Continuando con la búsqueda de un mejor modelo de GPR-54, se procedió a realizar un modelado ab initio mediante la plataforma trRosetta a la cual solamente se le brindó la secuencia en formato FASTA. El modelo obtenido (de TM-score: 0.81) se alineó por PYMOL con su template correspondiente (figura 13a) y se evaluó con ProSA, evidenciando un Z-Score dentro de los límites de distribución (figura 13b-d), y DALI con un rmsd de 0.8.



Figura 12. Alineamiento estructural de GPR-54 modelado y 4NH6 editada por PYMOL. En rojo se observa la estructura de GRR-54 y en verde la estructura de 4NH6. Rmsd = 1.3.

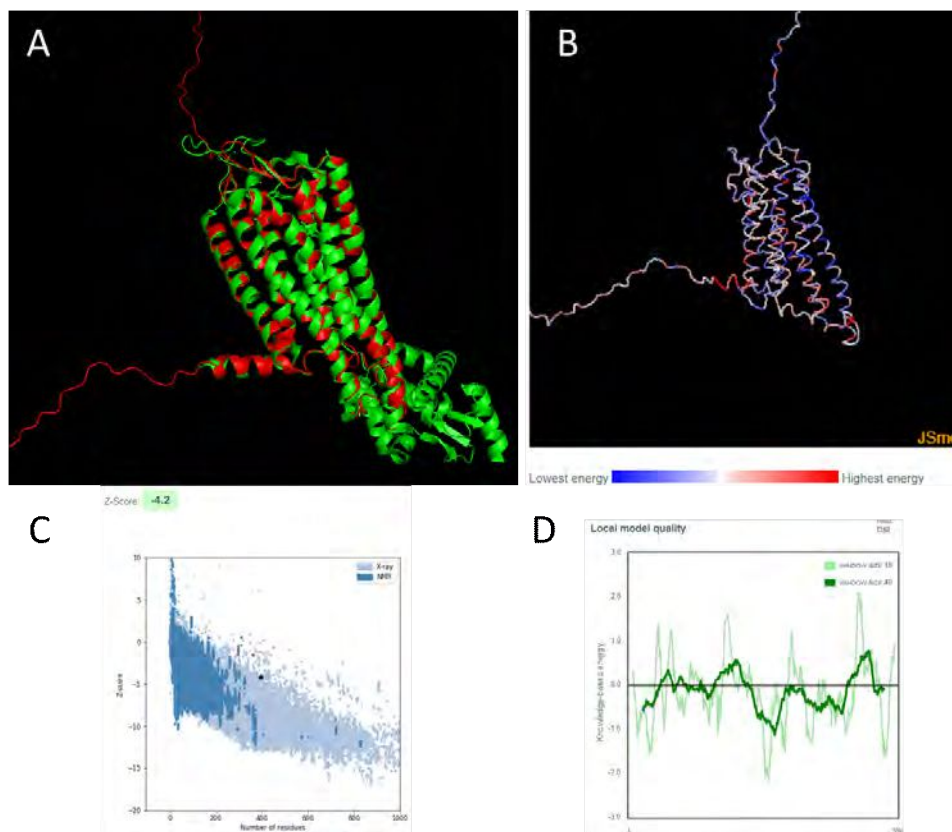


Figura 13. A. Alineamiento estructural por PYMOL de GPR-54 modelado ab initio y el template 5ZBH. En rojo se observa la estructura de GRR-54 y en verde la estructura de 5ZBH. Rmsd = 0.8, TM-Score = 0.81. B, C y D. Salida de ProSA para el modelo ab initio de GPR-54 de *Mus musculus*.

La asignación estructural se evaluó por la clasificación CATH, utilizando su servidor web. En este análisis a partir de la secuencia de Kiss-1 R de *Mus musculus*, se obtuvo como resultado que la proteína se compone de un solo dominio estructural, 4ea3B, que pertenece a la superfamilia de homólogos Rhodopsina de 7

hélices transmembrana. En la figura 14 se observa la salida completa del CATH con las 4 primeras categorías, clase, arquitectura, topología y superfamilia.

Level	CATH Code	Description
1		Mainly Alpha
1.20		Up-down Bundle
1.20.1070		Rhodopsin 7-helix transmembrane proteins
1.20.1070.10		Rhodopsin 7-helix transmembrane proteins

Figura 14. Salida del CATH para la secuencia del receptor Kiss-1 R de *Mus musculus*.
Clase: Mayormente alpha, Arquitectura: up-down Bundle, Topología y superfamilia de homólogos: Proteínas transmembrana de rodopsina de 7 hélices.

Una vez caracterizado el receptor de Kiss-1 secuencial y estructuralmente, se procedió al análisis de su filogenia utilizando las mismas secuencias del alineamiento múltiple y que destacaron la presencia de 6 grupos de proteínas homólogas con un patrón particular en la secuencia para cada una de ellas. Para ello, en una primera etapa se realizó la evaluación del modelo evolutivo a utilizar con el software Modeltest. En este caso se utilizó un árbol construido con el algoritmo de Neighbor Joining del programa HYPHY. Este árbol fue utilizado junto con el alineamiento múltiple previamente obtenido para la comparación de modelos vía Modeltest, con el establecimiento adicional de 4 categorías de velocidades de sustitución que toma la distribución gamma. Como resultado, el mejor modelo evolutivo para las secuencias y la topología dada, seleccionado según AIC, fue JTT + F.

A continuación se realizó la inferencia filogenética por Maximum Likelihood utilizando el programa PHYML (v. 3.1), utilizando como árbol de inicio un árbol de Neighbor Joining, y como algoritmo de búsqueda el mejor de NNI y SPR. Para el soporte de ramas se utilizó bootstrap, utilizando 100 replicantes.

El árbol resultante fue cargado en el servidor ITOL para su visualización y edición, obteniendo la figura 15 en la cual se puede apreciar las relaciones filogenéticas de Kiss-1 receptor isoforma 1 de *Mus musculus* y sus ortólogos, sin la presencia de root dada la falta de una secuencia outgroup.

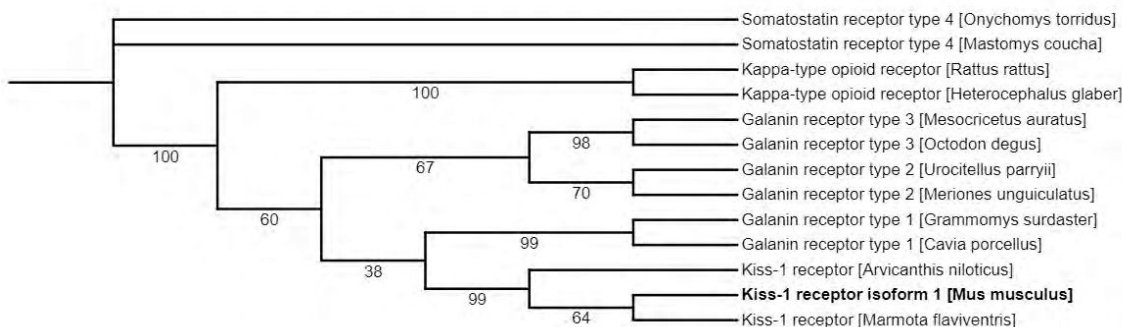


Figura 15. Árbol filogenético no roteado para el receptor de kiss-1 isoforma 1 de *Mus musculus* y los receptores homólogos. En cada rama se observa el soporte obtenido por bootstrapping.

El bootstrapping demuestra un sustento desde las secuencias muy robusto en el nodo entre los receptores de somatostatina y el resto de los ortólogos, lo mismo que sucede en relación al receptor opioide tipo Kappa. El soporte de la filogenia se encuentra en un rango de aproximadamente 65 en los nodos de los receptores de galanina, mientras que la información secuencial de estos con los Kiss-1 brinda un muy bajo soporte de la filogenia (38) en entre ambos.

Como último enfoque para la caracterización de GPR-54, se evaluó la conservación evolutiva y clase específica de aminoácidos que podría determinar la función proteica a partir de estudios con servidores

predictivos para dicho fin. El servidor ConSurf permitió establecer cuáles son los aminoácidos conservados y los mapeo en la estructura template 4N6H utilizada para el modelado por homología, destacando una amplia conservación en las regiones alfa hélice pertenecientes a los segmentos transmembrana.

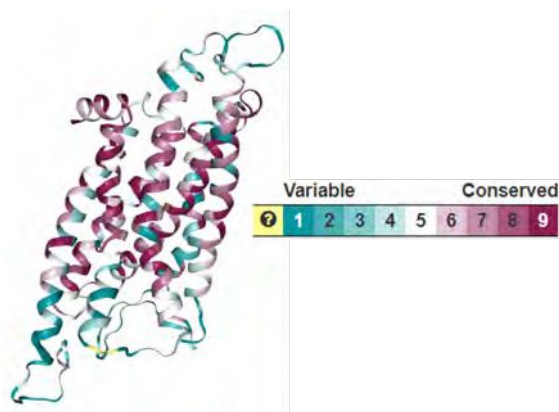


Figura 16. Visualización de la conservación aminoacídica mapeada en la estructura del modelo creado a partir del template 4N6H, mediante el servidor ConSurf.

Al aplicar el método de análisis del Evolutionary Trace (ET) al alineamiento múltiple de secuencia, se encontraron aproximadamente 20 aminoácidos de importancia evolutiva clase específica (figura 17), centrados también en la región transmembrana y no haciéndose presentes en los extremos amino y carboxilo terminal.



Figura 17. Secuencia de GPR-54 de *Mus musculus* con la importancia evolutiva por posición según el servidor Evolutionary Trace.

Cabe mencionar que en el análisis ET clase específico, los aminoácidos L102, L148, C223 cuya mutación se asocia a patologías por pérdida de función, no son reflejados como de importancia evolutiva funcional a la vez que el residuo R331 si es de relevancia destacándose en naranja en la secuencia de la figura 18.

CONCLUSIONES Y DISCUSIÓN

A partir de los métodos bioinformáticos utilizados se logró cumplir con los objetivos establecidos de determinar la máxima distancia evolutiva mediante homólogos del receptor GPR-54, encontrando secuencias de hasta un 20% de identidad en especies tan alejadas evolutivamente de *Mus musculus* como ser *Drosophila melanogaster*, que por análisis de dominios se relacionaron mediante la superfamilia 7tm_GPCRs.

Además se logró establecer las diferencias secuenciales presentes entre las macromoléculas homólogas en los subórdenes *Myomorpha*, *Hystricomorpha* y *Sciuromorpha*, destacando la presencia de 6 grupos de

proteínas homólogas las cuales comparten una gran cantidad de posiciones completamente conservadas, una mayor proporción de aminoácidos de carácter hidrofóbico y mutaciones sitio específicas que se comparten entre las proteínas de un grupo en las regiones transmembrana caracterizadas, a la vez que los segmentos de loop fueron de extensión grupo de proteínas específico. A nivel de la comparación de secuencias también se pudo corroborar la conservación de información de las posiciones aminoacídicas 102, 148, 223 y 331, cuya mutación deriva en infertilidad.

Los diversos predictores de estructura secundaria permitieron caracterizar exitosamente los elementos Hélices α , hoja β y coiled coils, además de los segmentos transmembrana y las regiones con desorden de la secuencia del receptor GPR-54. Todas estas estructuras fueron visualizadas espacialmente a partir del modelado por homología y ab initio, de RMSD = 1.3 y 0.8, respectivamente.

Por último en la caracterización de las posiciones evolutivamente relevantes para la función de GPR-54, cabe destacar la falta de coincidencia entre residuos que se encontraron conservados y que por bibliografía son importantes, dado que su mutación lleva a la pérdida funcional, y los residuos clase específicos, hipotetizando que su valía vendría de una relación epistática con aminoácidos de cercanía que generan un ambiente fisicoquímico particular y que mostraron estar conservados y ser funcionalmente importantes.

A modo de conclusión se logró caracterizar secuencial, estructural y evolutivamente a GPR-54, en una triada que se interrelaciona desde la perspectiva estructura-función y que es el reflejo del paso evolutivo.

BIBLIOGRAFÍA

Civelli O, Zhou QY (2008) Orphan G Protein-Coupled Receptors and Novel Neuropeptides. 10.1007/400_2007_050

Clarkson J, d'Anglemont de Tassigny X, Moreno AS, Colledge WH, Herbison AE. (2008). Kisspeptin-GPR54 signaling is essential for preovulatory gonadotropin releasing hormone neuron activation and the luteinizing hormone surge. *Journal of Neuroscience*. 28:8691–8697.

Clements MK, McDonald TP, Wang R, Xie G, O'Dowd BF, George SR. (2001) FMRamide-related neuropeptides are agonists of the orphan G-protein-coupled receptor GPR54. *Biochem Biophys Res Commun*; 284:1189–93.

Evans BJ, Wang Z, Mobley LL, Khosravi D, Fujii N, Navenot JM, Peiper SC: (2008) Physical association of GPR54 C-terminal with protein phosphatase 2A. *Biochem Biophys Res Commun*; 377:1067–1071.

Kauffman AS, Clifton DK, Steiner RA. (2007). Emerging ideas about kisspeptin- GPR54 signaling in the neuroendocrine regulation of reproduction. *Trends in Neurosciences*. 30(10): 504–511.

Kotani M, Dethoux M, Vandenberghe A, Communi D, Vanderwinden JM, Le Poul E. (2001) The metastasis suppressor gene *KiSS-1* encodes kisspeptins, the natural ligands of the orphan G protein-coupled receptor GPR54. *J Biol Chem*; 276:34631–6.

Lee DK, Nguyen T, O'Neill GP, Cheng R, Liu Y, Howard AD. (1999) Discovery of a receptor related to the galanin receptors. *FEBS Lett*; 446:103–7.

Lehman MN, Coolen LM, Goodman RL. (2010). Minireview: kisspeptin/neurokinin B/dynorphin (KNDy) cells of the arcuate nucleus: a central node in the control of GnRH secretion. *Endocrinology*. 151: 3479–3489.

Messager S, Chatzidakis EE, Ma D, Hendrick AG, Zahn D. (2005). Kisspeptin directly stimulates gonadotropin-releasing hormone release via G protein-coupled receptor 54. *Proceedings of the National Academy of Sciences of the United States of America*. 102(5): 1761–1766.

Muir AI, Chamberlain L, Elshourbagy NA, Michalovich D, Moore DJ. (2001). AXOR12, a novel human G protein-coupled receptor, activated by the peptide *KiSS-1*. *Journal of Biological Chemistry*. 276(31):28969–28975.

Navarro VM, Castellano JM, Fernandez-Fernandez R, Tovar S, Roa J. (2005). Effects of *KiSS-1* peptide, the natural ligand of GPR54, on follicle-stimulating hormone secretion in the rat. *Endocrinology*. 146(4): 1689–1697.

Ohtaki T, Shintani Y, Honda S, Matsumoto H, Hori A, Kanehashi K, Terao Y, Kumano S, Takatsu Y, Masuda Y. (2001). Metastasis suppressor gene *KISS-1* encodes peptide ligand of a G-protein-coupled receptor. *Nature*. 411: 613–617.

Pasquier J, Kamech N, Lafont AG, Vaudry H, Rousseau K, Dufour S. (2014) Molecular evolution of GPCRs: Kisspeptin/kisspeptin receptors. *J Mol Endocrinol*; 52: 101-117.

Seminara SB, Messager S, Chatzidaki EE, Thresher RR, Acierno JS, Shagoury JK, Bo-Abbas Y, Kuohung W, Schwinof KM, Hendrick AG, Zahn D, Dixon J, Kaiser UB, Slaughter SA, Gusella JF, O’Rahilly S, Carlton MB, Crowley WF Jr, Aparicio SA. (2003) The *GPR54* Gene as a Regulator of Puberty. *N Engl J Med*; 349:1614–1627.

Conclusiones

En este libro se presentaron los trabajos desarrollados por los alumnos de la asignatura Bioinformática de la Facultad de Ciencias Exactas de la Universidad de La Plata. Dichos alumnos representan dos instancias diferenciadas de la formación académica, incluyendo alumnos de grado de los últimos años de la carrera de Licenciatura en biotecnología y biología molecular (en mayor proporción), y de postgrado, en general, de los primeros años de diferentes carreras de doctorado de la UNLP. Estos trabajos se presentaron como requisito para la aprobación de la parte práctica de la materia, y con el objetivo principal de formalizar los conocimientos adquiridos por los alumnos a lo largo de la cursada.

En años anteriores, la evaluación de los conocimientos prácticos era realizada mediante un examen en el cual los alumnos debían responder una serie de preguntas utilizando las aplicaciones bioinformáticas desarrolladas en los sucesivos trabajos prácticos. Si bien la mayoría de los alumnos era capaz de responder correctamente las preguntas durante la evaluación práctica, las situaciones que se encuentran al trabajar en bioinformática suelen presentar desafíos para los cuales los alumnos no estaban preparados. Ante esta situación, surgió la idea de que los alumnos caracterizaran bioinformáticamente una proteína de su elección (en el caso de alumnos de postgrado), o proporcionada por la cátedra, de forma simultánea a la realización de los diferentes trabajos prácticos y que confeccionaran un informe detallado con formato de publicación. En este contexto cada alumno recibió una secuencia aminoacídica a partir de la cual confeccionaron los informes publicados en este libro.

Como se describió en la presentación, los trabajos presentados debían contener una serie de estudios bioinformáticos como la búsqueda de homólogos cercanos y remotos, la creación de un alineamiento múltiple y la inferencia filogenética a partir del mismo. Finalmente, los alumnos debían obtener un modelo molecular por el método de modelado por homología. Vale aclarar, que durante el comienzo del año 2021 todavía no existía la posibilidad de obtener modelados *ab initio* mediante AlphaFold2 (y mucho menos la base de datos AlphaFold DB), método que ha desbancado al modelado por homología como estándar para la obtención de estructuras de proteínas.

A lo largo del desarrollo de este trabajo los alumnos se vieron enfrentados a multitud de problemas particulares que debieron resolver con asistencia de la cátedra a través de consultas personales. Ejemplos del tipo de problema que debieron resolver incluyó la utilización de diferentes bases de datos secuenciales o la aplicación de filtros en las búsquedas, la selección de secuencias a utilizar para la generación de alineamientos múltiples y estudios filogenéticos con un objetivo particular, y resolver los problemas particulares de cada secuencia para la obtención de un modelo molecular con Modeller. Finalmente, se valoró el análisis integral de los resultados obtenidos por las diferentes metodologías.

Otra observación realizada por la cátedra fue que los alumnos presentaban dificultades a la hora de expresar los conocimientos adquiridos, tanto en forma oral como escrita. Mediante la presentación de un trabajo redactado con lenguaje técnico y de forma estructurada, pretendimos mejorar la calidad de expresión de los alumnos, señalando los errores y sugiriendo formas de mejorar la redacción.

Dicho esto, estamos más que satisfechos con los resultados obtenidos. Como puede observarse en los trabajos presentados, los objetivos planteados se cumplieron holgadamente, presentando la comunicación de los resultados obtenidos con un lenguaje adecuado y de forma concisa. Los resultados obtenidos de las diferentes aplicaciones bioinformáticas utilizadas fueron correctamente analizados, y en ocasiones, se obtuvieron conclusiones interesantes sobre las proteínas en estudio.

Nos consta que el desarrollo de este trabajo ha contribuido favorablemente a la formación de los alumnos, tanto de los que optaron por publicar sus resultados en este libro como de los que no. El desarrollo de estos trabajos requirió un compromiso que fue más allá de la aprobación de la materia, y que en muchos casos implicó profundizar un poco más los estudios realizados para la confección de este libro. Agradecemos a los alumnos por su dedicación y esperamos haber dejado una impronta en su formación académica.

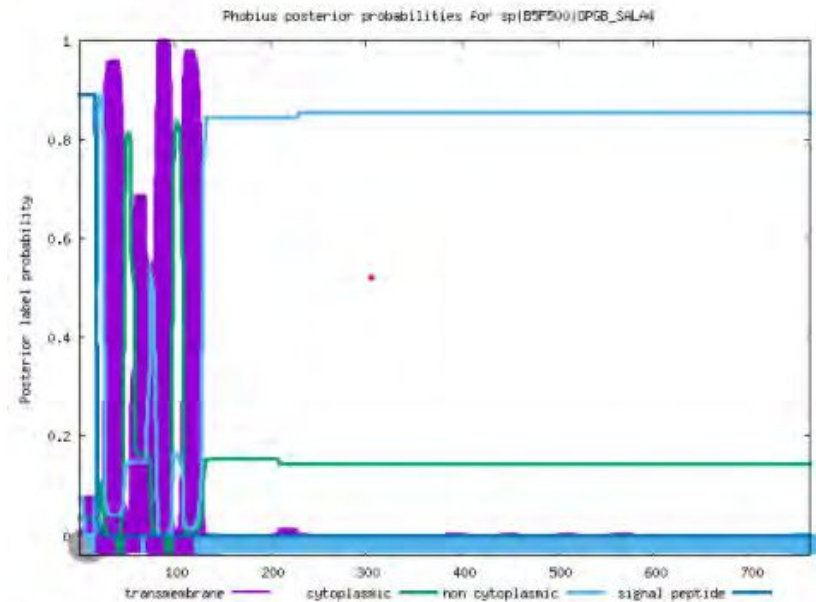
Apendice

A1 - Material suplementario: Análisis secuencial de la proteína fosfoglicerol transferasa I de Salmonella agona (cepa SL 483)

Phobius prediction

Prediction of sp|B5F500|OPGB_SALA4

ID	sp B5F500 OPGB_SALA4		
FT	SIGNAL	1	18
FT	REGION	1	3
FT	REGION	4	14
FT	REGION	15	18
FT	TOPO_DOM	19	27
FT	TRANSMEM	28	47
FT	TOPO_DOM	48	58
FT	TRANSMEM	59	74
FT	TOPO_DOM	75	79
FT	TRANSMEM	80	98
FT	TOPO_DOM	99	109
FT	TRANSMEM	110	128
FT	TOPO_DOM	129	763
FT	//		



Predicción de estructura secundaria a partir de secuencia de aminoácidos con el programa Phobius.

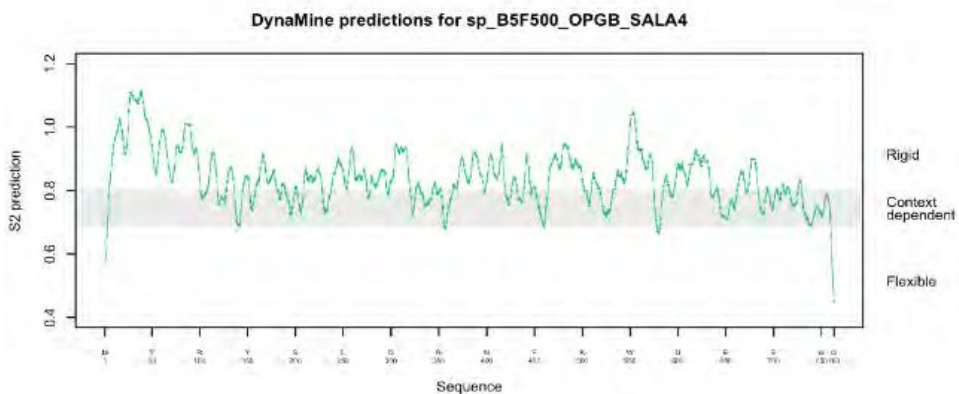


Fig S1. Análisis del desorden de la enzima fosfoglicerol transferasa I mediante el programa Dynamine.

Tabla S1. Código de acceso de las secuencias seleccionadas para el AM y sus organismos asociados.

Código UniProt	Taxonomía
B5F500 OPGB_SALA4	<i>Salmonella agona (strain SL483)</i>
OPGB_SALTY	<i>Salmonella typhim</i>
A0A2X4TXM2_SALER	<i>Salmonella enterica</i>
A0A0V9JJ53_9ENTR	<i>Citrobacter sp.</i>
OPGB_ECOLI	<i>Escherichia coli (strain K12)</i>
OPGB_SHIFL	<i>Shigella flexneri</i>
Q327N7_SHIDS	<i>Shigella dysenteriae</i>
D2TRV2_CITRI	<i>Cyrobacter rodentium</i>
LOMAT2_ENTBF	<i>Enterobacteriaceae bacterium (strain FGI57)</i>
A0A2X3ECK0_KLUCR	<i>Kluyvera cryocrescens</i>
A0A0H3GM60_KLEPH	<i>Klebsiella pneumoniae</i>
E3GAV4_ENTLS	<i>Enterobacter lyngnoliticus</i>
A0A2P5GQK4_9ENTR	<i>Superficieibacter electus</i>
A0A2X2E7L4_RAOPL	<i>Raoultella planticola</i>
A0A085AB46_9ENTR	<i>Trabulsiella guamensis</i>
A0A0J8VPZ9_9ENTR	<i>Franconibacter pulveris</i>
A0A2P8VMR8_9ENTR	<i>Siccibacter turicensis</i>
I2BD48_SHIBC	<i>Shimwellia blattae</i>
A0A071LXN8_9ENTR	<i>Mangrovibacter sp.</i>
A0A4P8YID1_9ENTR	<i>Izhakiella sp.</i>
A0A085GIE3_9ENTR	<i>Buttiauxella agrestis</i>
A0A1T4P151_ENTAG	<i>Enterobacter agglomerans</i>
A0A0L7SYL9_9GAMM	<i>Erwinia iniecta</i>
A0A1X1CX21_9GAMM	<i>Pantoea wallisii</i>
A0A4C1SNN7_9NEOP	<i>Eumeta japonica</i>
A0A2P5IV60_9GAMM	<i>Pantoea sp. PSNIH6</i>
Código UniProt	Taxonomía
B2VL27_ERWT9	<i>Erwinia tasmaniensis</i>
A0A085JKA2_9GAMM	<i>Tatumella tyseos</i>
A0A240C7R6_SERFI	<i>Serratia ficaria</i>
H2ISE6_RAHA6	<i>Rahnella aquatilis</i>
A0A1I4VV11_9ENTR	<i>Izhakiella capsodis</i>
A0A085GIW9_9GAMM	<i>Ewingella americana</i>

A2 - Material suplementario: Análisis bioinformático de una cisteín proteasa presente en plantas de *Nicotiana benthamiana*, la proteína Metacaspasa 4.

Figuras Suplementarias

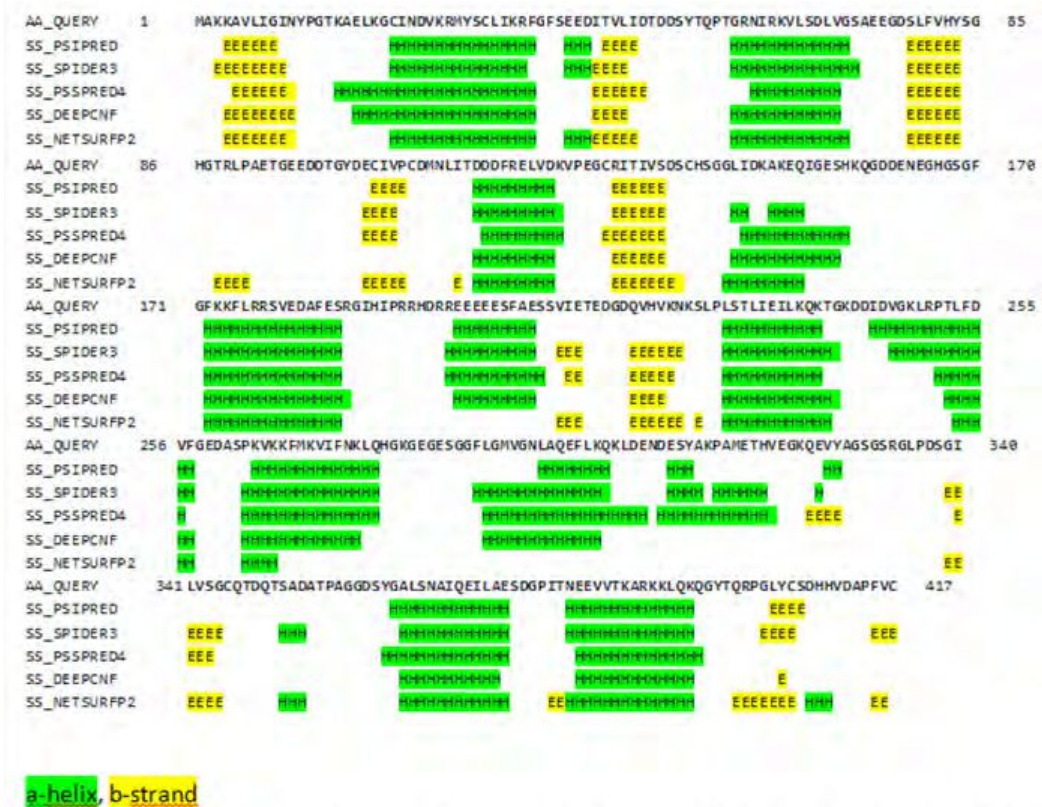


Figura S1. Predicción de estructuras secundarias a partir de la secuencia primaria a partir de la plataforma Quick2d. Las referencias por colores se observan en el extremo inferior.

10 20 30 40 50 60 70 80 90

Achinense/1-430 MRKERNRREREERE...GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

A.officinale_MCA4/1-417GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

A.precatorius_MCA5-like/1-420GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

A.rufa_MCA5/1-421GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

A.yangbiense_hypothetical/1-419GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

B.hispida_MCA4/1-418GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

Cannum_MCA4/1-416GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

Carabica_MCA4/1-416GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

Caryopema_subsp.aryzopema_MCA4/1-432GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

C.orientalis_hypothetical/1-429GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

C.baccatum_MCA4/1-417GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

C.angustata_unnamed/1-416GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

C.fingiana_hypothetical/1-419GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

C.giloinensis_hypothetical/1-417GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

C.giloinensis_hypothetical/1-417GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

C.mexica_MCA4/1-424GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

C.melo_var.maluwa_MCA4/1-422GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

C.mochata_MCA4/1-424GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

C.papaya_MCA5/1-417GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

C.pepo_subsp.pepo_MCA4/1-424GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

C.sativus_MCA5/1-422GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

C.sinenis_MCA5/1-422GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

D.orientalis_subsp.sativus_MCA4/1-416GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

D.zibethinus_MCA4/1-415GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

C.orientalis_unnamed/1-429GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

E.guttata_MCA4/1-415GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

H.braziliensis_MCA4/1-418GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

C.cepulans_Peptidase/1-417GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

H.impetiginosus_hypothetical/1-435GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

Humboldtia_MCA5/1-414GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

Ithoba_MCA4/1-419GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

Jouana_MCA5/1-419GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

J.licocarpa_X/regia_MCA4/1-416GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

J.regia_MCA4/1-416GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

K/1-416GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

Lalibus_sutaiwei/1-419GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

Languatfolius_MCA5/1-419GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

L.chinensis_MCA4/1-418GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

M.orientalis_Peptidase/1-420GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

M.notalia_MCA4/1-433GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

M.truncatula_MCA5/1-413GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

M.tenuata_MCA4/1-418GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

M.benthiana_MCA4/1-418GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

N.sinenis_hypothetical/1-425GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

N.sylvestris_MCA4/1-418GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

N.tabacum_MCA4/1-418GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

N.tomentosifomis_MCA4/1-418GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

O.orientalis_subsp.europaea_Hypothetical/1-423GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

O.orientalis_var.sylvestris_MCA4/1-419GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

P.alba_MCA4/1-421GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

P.alba_MCA5/1-420GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

P.andersonii_Caspae-like/1-431GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

P.deltoides_hypothetical/1-425GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

P.euphratica_MCA4/1-422GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

P.zoaniferum_MCA4/1-407GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

P.tomentosa_hypothetical/1-424GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

P.trichocarpa_MCA4/1-423GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

P.vera_MCA4/1-417GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

PetuniaXhybrida_MCA2/1-417GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

R.argentea_MCA4/1-419GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

R.communis_MCA4/1-420GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

R.sinal_hypothetical/1-438GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

S.braehista_hypothetical/1-423GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

S.commesoni_hypothetical/1-417GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

S.denticosa_Peptidase/1-417GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

S.dunni_hypothetical/1-432GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

S.indicum_MCA4/1-422GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

S.lyopersicum_MCA1/1-417GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

S.pennellii_MCA4/1-417GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

S.sutchowensis_MCA4/1-417GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

S.tuberosum_MCA4/1-417GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

T.cacao_MCA4/1-414GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

T.pratense_MCA4/1-413GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

T.sinenis_hypothetical/1-418GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

T.wilfordii_MCA4/1-418GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

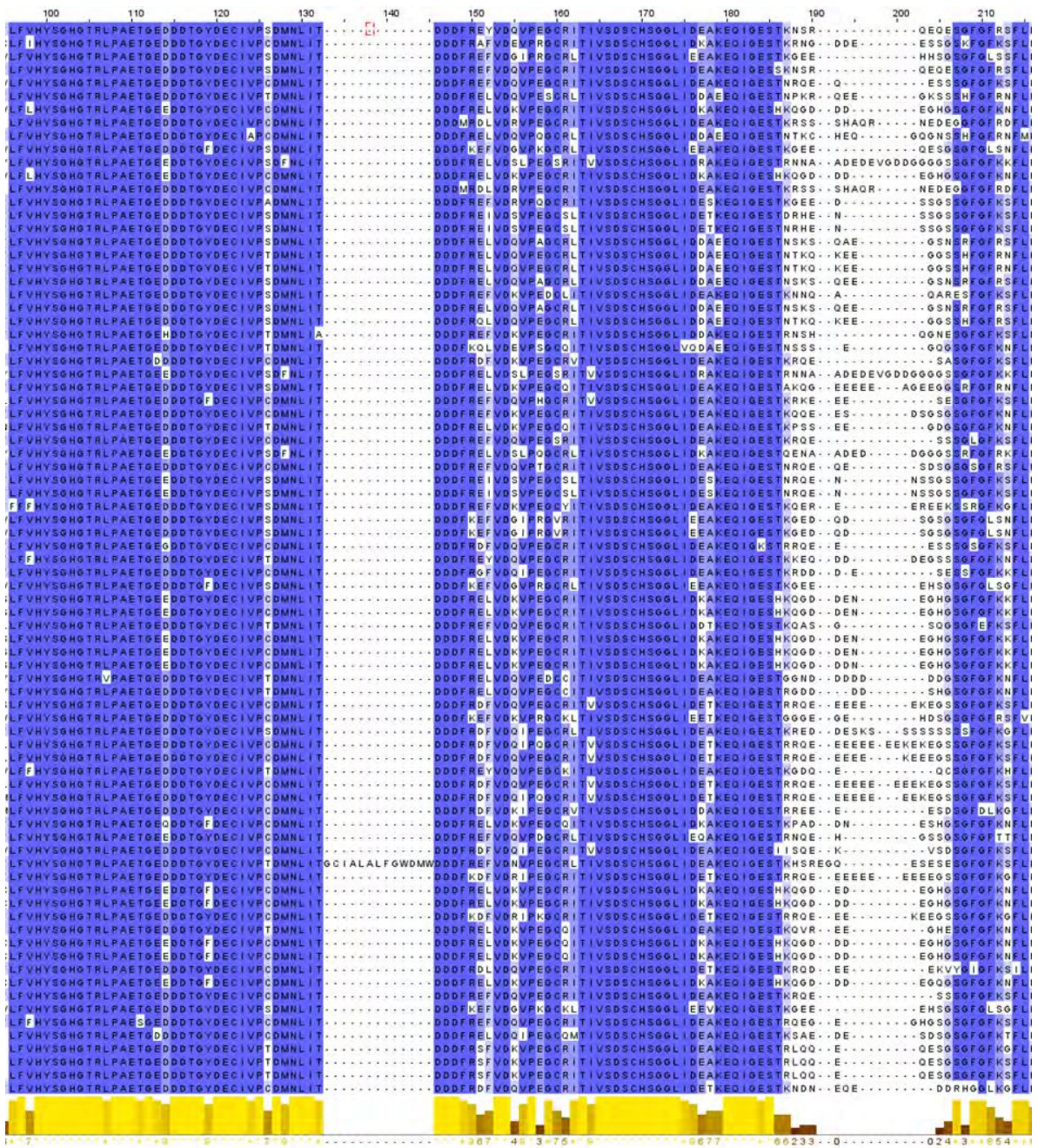
V.xiphioides_MCA5/1-425GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

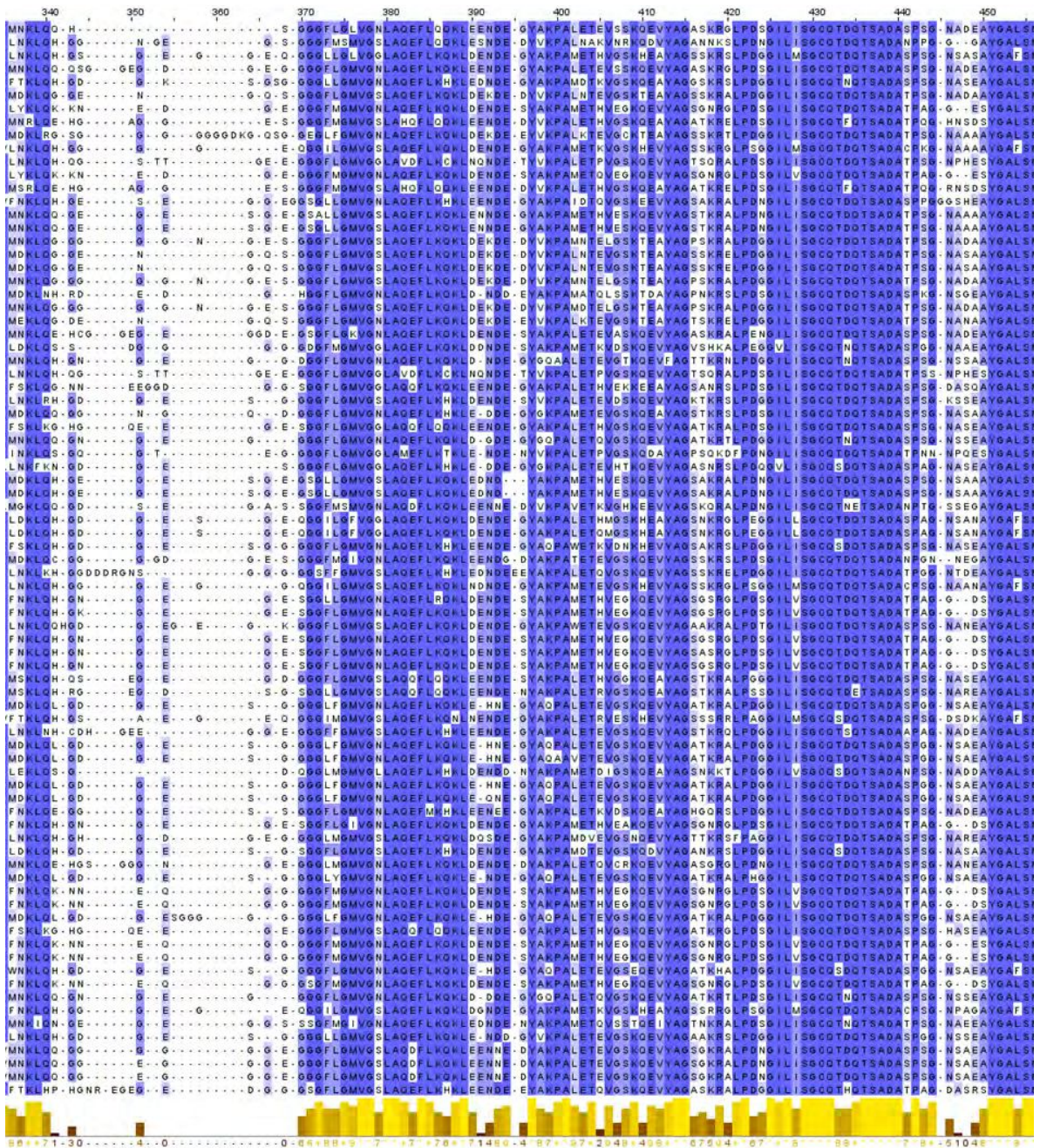
V.ripestis_MCA5/1-425GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

V.vinifera_MCA5/1-425GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF

Z.jubba_MCA5/1-423GKAVLLIGINYPGKAEKGGVNDVKRMYS...DLVERVGRFEDITVLDITDDEYVTPGTHHRRKALSDLLRSAPGDF







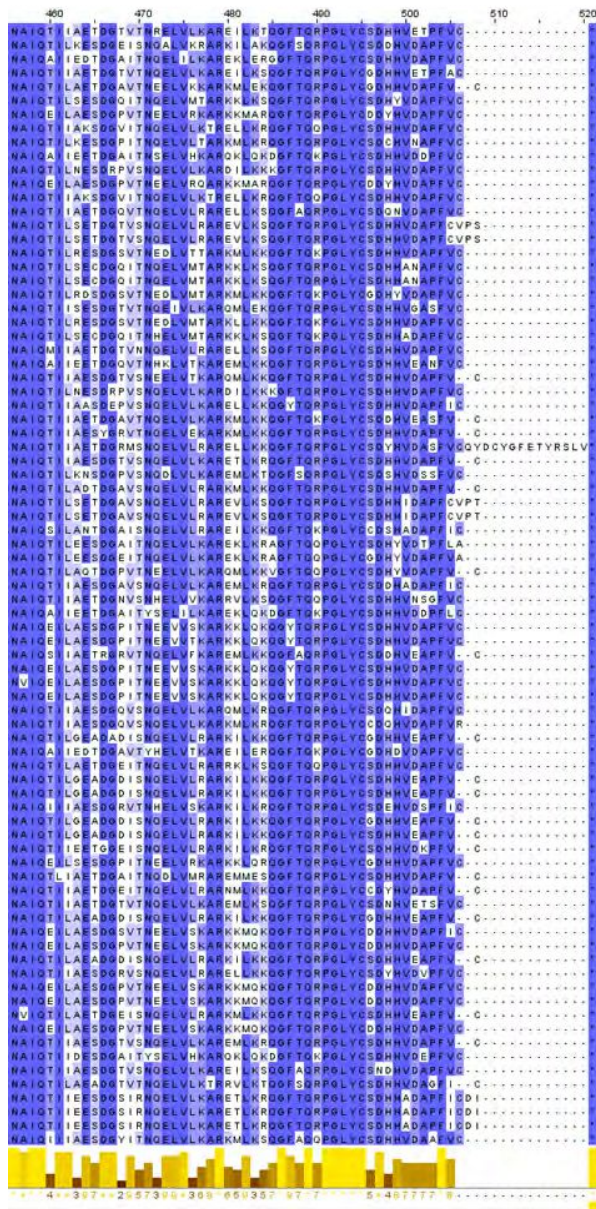


Figura S2. Alineamiento múltiple entre NbMCA4 y 81 homólogas. El gráfico amarillo indica el nivel de conservación de los aminoácidos, al igual que la mayor intensidad en el color azul. Se indica el género y especie de cada proteína.

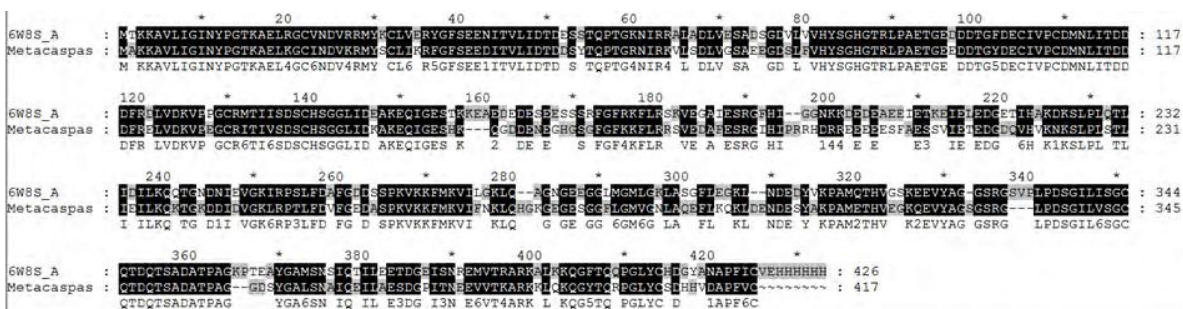


Figura S3. Alineamiento de a pares entre NbMCA4 y su homóloga con estructura conocida, AtMCA4 (Código PDB 6W8S).

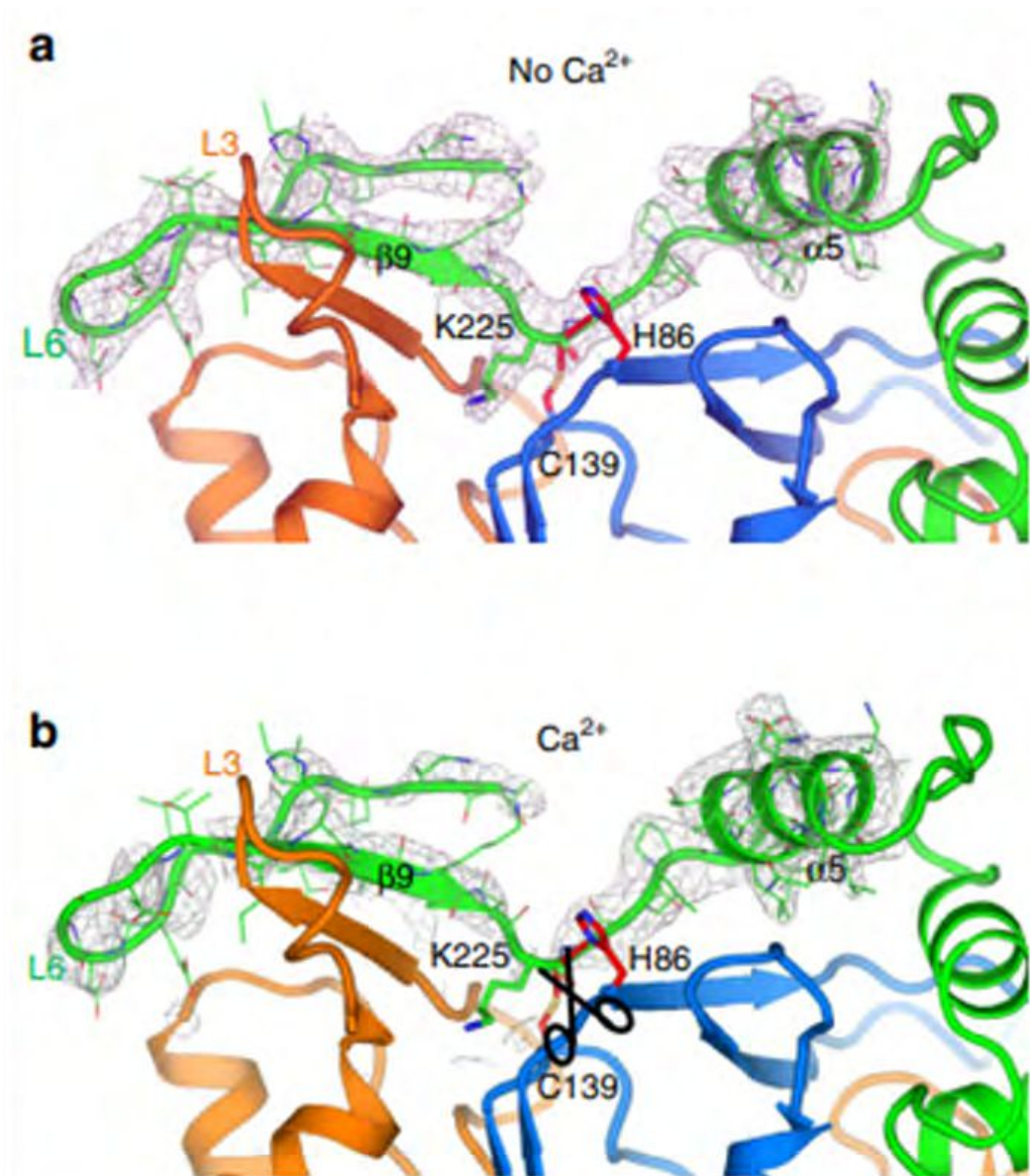


Figura S4. Estructura cristalizada de la proteína AtMCA4 con los residuos del sitio activo (H86 y C139) y el sitio de clivaje (K225) marcados. a. Proteína no tratada. B. Microcristales de la proteína tratada con Ca²⁺, activando el clivaje en K225. Imagen original tomada de Zhu et al. 2020.

A3 - Material suplementario: Modelado molecular y predicción funcional de la proteína bli4781 de *Bradyrhizobium diazoefficiens* (USDA 110)

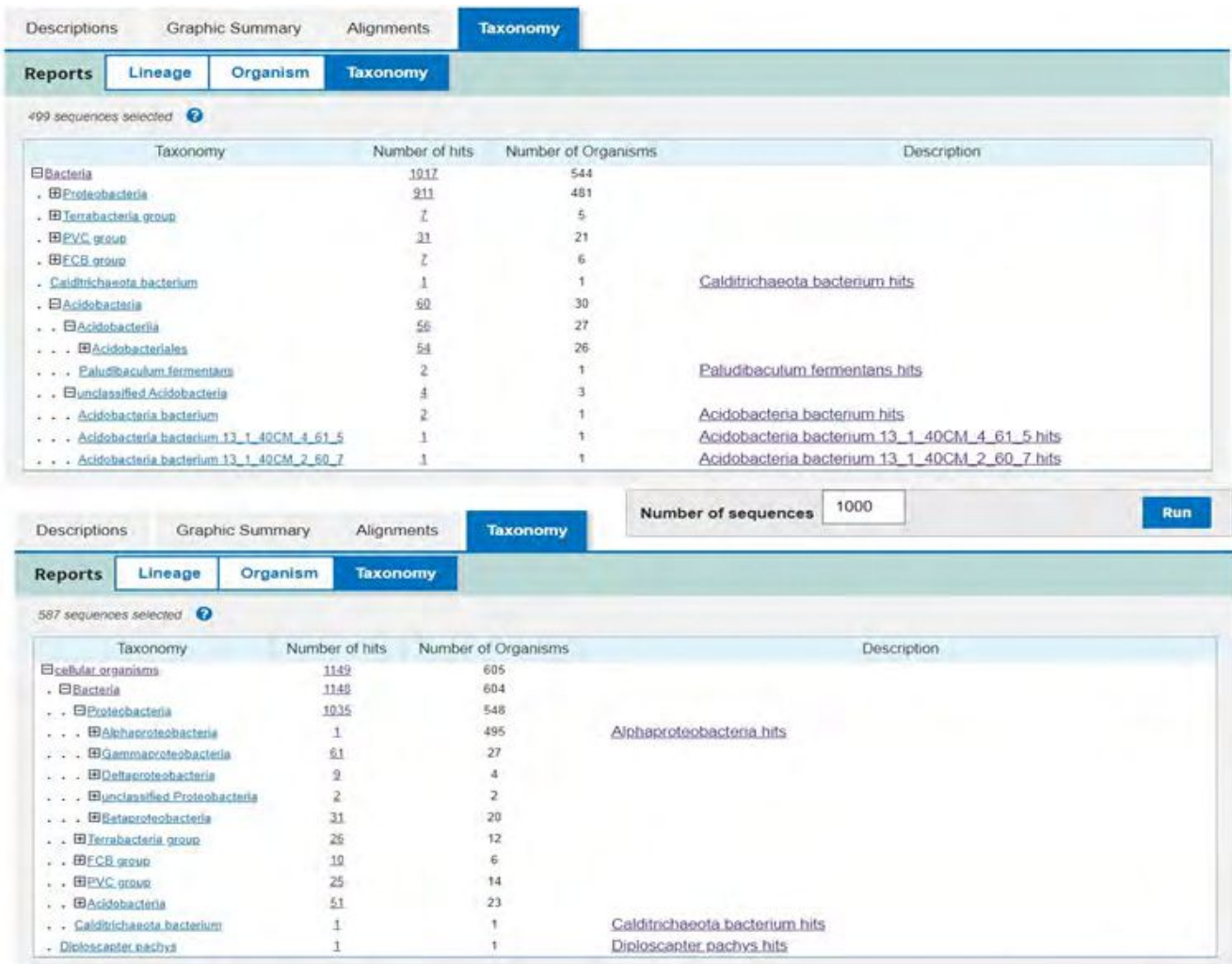


Figura S1. Distribución taxonómica en BLAST (arriba) y PSI-BLAST (abajo)

Tabla S1. Secuencias elegidas para realizar el alineamiento múltiple

DUF5076 (Especies)	Accession number	Secuencia
<i>Bradyrhizobium erythrophlei</i>	WP_079603347.1	MSGPKQPLPPDVIGRDDATEVLRAFIVDGGLSIAFTRAFEEPMWGLLVDL ARHAARAYARESAFTEDEALTRIVDMFEAEIARPTDPGTTTPRSQQGH
<i>Bradyrhizobium elkani</i>	GEC52187.1	MAGPKQPLPPDVMGRDDAIEVLRAFVVDGGLSIAFQRAFEPMWGLML VDIARHAARAYARESEYTEDEALARIVEMFEAEIARPTDMGQTKPRSQQGH
<i>Bradyrhizobium yuanmingense</i>	TGN80551.1	MAGPKQPLPPDVMTRDDAVEILRVFVLDGGLSMAFQRAFEPMWGLLL VDLARHAARAYARESEYTEEDALNRILDMFQAEIERPTDTGTTTPRGKGH
<i>Bradyrhizobium</i> sp. Tv2a-2	WP_024518800.1	MAGPKQPLPPDVIGREDATEVLRAFVLDGGLSIAFTRAFEEPMWGLLVDI ARHAARAYSRESYSEEEALDRILEMFSAEIERPTDMGTTTPRSQKGH
<i>Rhodopseudomonas palustris</i>	WP_119857868.1	MAGPKQPLPPDVEGREDATEVLRAFVLDGGLSIAFMRAFEPEMWWGLLV DIARHAARAYSRESNYTEDEALERIVEMFESEIARPSDLGSTTERPQ
unclassified <i>Tardiphaga</i>	WP_143573479.1	MAGPKQPLPPDVVGREEAVEVLRAFVVDGGLSIAFTRAFEEPMWGLLV DIARHAARAYARESDYSEDEALARIVDMFDAEIERPTDVGNTTTPRSQQGH
<i>Microvirga aerophila</i>	WP_114188594.1	MTEKKFEALNVPPDVLEKGGVEILRASVVDGAVSIALRRAFDDPFTWGVLLVD LARHAARVYAMETDFSEEEALAEISAGIQAEILDDPDPGPTQAIN
DUF5076 (Especies)	Accession number	Secuencia
<i>Methylobacterium</i> sp. YIM 132548	WP_150964006.1	MPKAFQPLSVPDALEKGGVEVLRASVVDGAVSVALRRSFDPPFTWGVLLID LARHAARVYAEITDLSEEEFAFIQIRAGLEAETDPPDGPDSLLN

<i>Mesorhizobium sp.</i> WSM4312	PBB66311.1	MFGKLSSELSPPPNAKNARAVEVLRVWAEPGAQQLVKTTWKPEGAWGLL LADVARHAARAYVAEGISEAQALDRILMLFKAFAAPTDPAPS
<i>Sphingomonas sp.</i> Leaf230	WP_056057410.1	MADHPANHPNAIALDKGAQLTGESVEVARIWITNGAGSNVLDAGILEDPYTF GYLLADTIRHAARAYAGTWGLDEDAALQAIVDGVGTLELREQFTTITTIQEGMH
<i>Dechloromonas sp.</i> CZR5	WP_150427584.1	MFGRKGIELEPPLSRDAGAFEILRVWGGDNLPQQYSLKTVRDDPGAWGLM LVDIARHVAKAYGNTGDFSEEAALKRIKELFDAEWASPTDTPQLVK
<i>Ralstonia sp.</i>	WP_048934697.1	MKSELPIPPAALQDVDSKELLRVWAAAGNQHSIATGVWENPTIWIWIMLVDL LRHIARSYKVGNVSYEESMRLIKAGFDAEWESPT
<i>Lactobacillus crispatus</i>	WP_133467175.1	MAGPKQLPDPVVMGRDDAVEVLRVFDGGLSIAFQRAFEEDPMWGLM LVDIARHAARAYSRESEYTEDEALARIVEMFEAEIARPTD
<i>Granulicella sp.</i> S156	WP_158821490.1	MSGNKYLDPPAAVRDKASFELLRVWVAEQGQHVSLRPGTWDDPFAWGV LADLARHIVNAESIHRKNFDEDAFLERMLLEGFRAIESPTDDPEGEIMQ
<i>Calditrichaeta bacterium</i>	RMI08682.1	MAAENPDAFEVLRWITAPGAYQVILRTSWEDPGAWGILLVDIARHAARAYE REGWDRREALDRIRELDAEWDFPTDEPLDITRDS



The conservation scale:



Variable Average Conserved

- e - An exposed residue according to the neural-network algorithm.
- b - A buried residue according to the neural-network algorithm.
- f - A predicted functional residue (highly conserved and exposed).
- s - A predicted structural residue (highly conserved and buried).

Figura S2. En escala de colores, posiciones menos y más conservadas según el programa ConSurf.

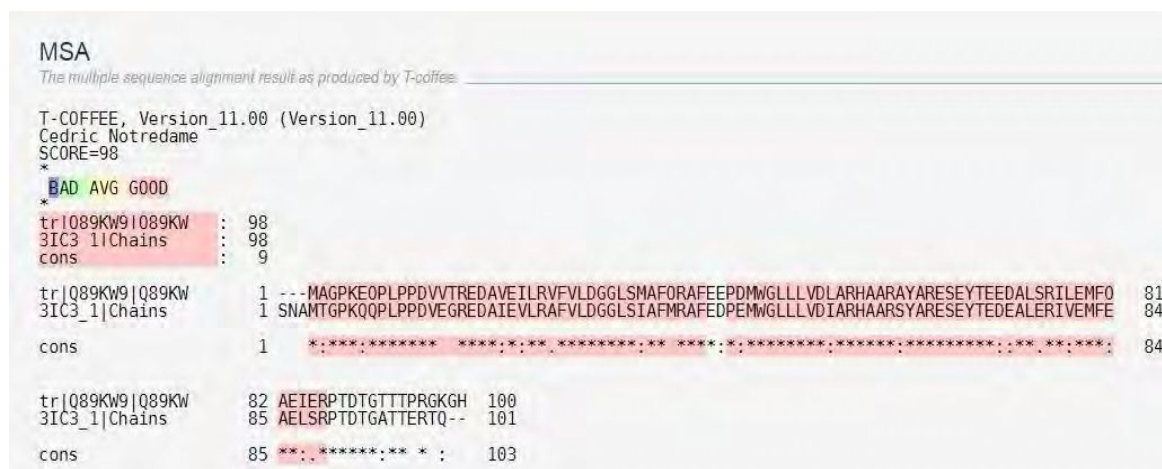
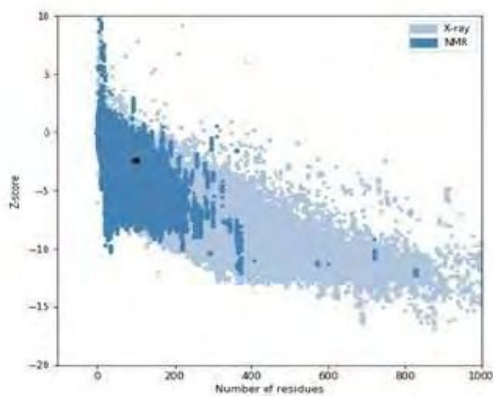


Figura S3. Alineamiento entre la query sequence y la template generado por T-Coffee.

Overall model quality

Z-Score: -2.43



Local model quality

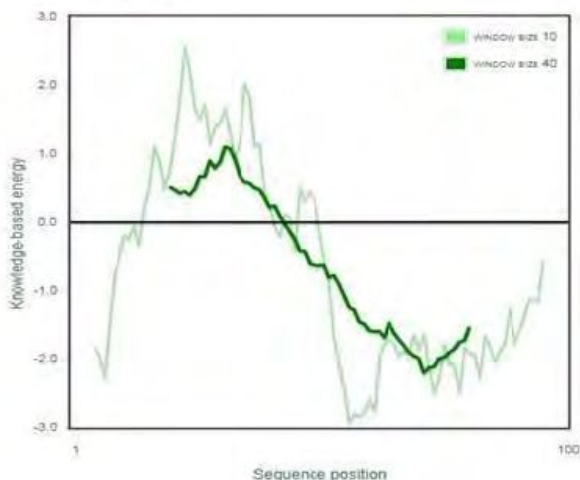


Figura S4. Resultados gráficos de evaluación y perfil energético obtenido por ProSa.

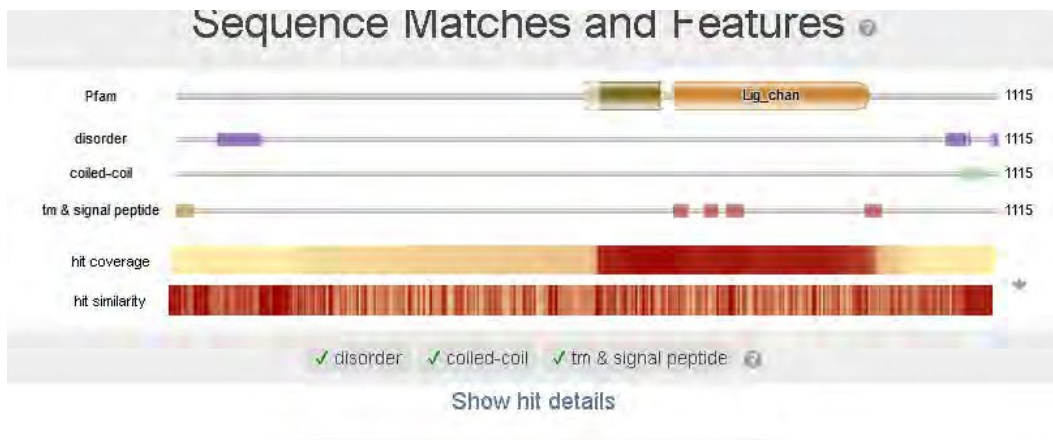
```

tr|Q89KW9| MAGPK--EQP L-----PPDV VTREDAVEIL RVFVLDGGL- SMAFQ-RAFE EPDMGMLLV DLARHAARAY A-RESEYTEE DALSRILEMF QAEIERPTDT GTTTPRGKG- H
WP_0796033 MSGPK--EQP L-----PPDV IGRDDATEVL RAFIVDGG- SIAFT-RAFE EPDMGMLLV DLARHAARAY A-RESAPTED EALTRIVDMF EAEIARPTDP GTTTPRSQQG H
GEC52187.1 MAGPK--EQP L-----PPDV MGRDDAIEVL RAFVVDGGL- SIAFQ-RAFE EPDMGMLLV DLARHAARAY A-RESEYTED EALARIVEMF EAEIARPTDM GQTKPRSQQG H
RMI08682.1 -----MA- AENPDAFEVL RIWTAPGAV- QQVILRTSWE DPGAWGILLV DIARHAARAY E-REG-WDRR EALDRIRELF DAEWDFPTDE PLDITR-DS- -
WP_0245188 MAGPK--EQP L-----PPDV IGRDDATEVL RAFVLDGGL- SIAFT-RAFE EPDMGMLLV DLARHAARAY S-RESDYSEE EALDRILEMF SAELERPTDM GTTTPRSQKG H
WP_1198578 MAGPK--QQP L-----PPDV EGREDATEVL RAFVLDGGL- SIAFM-RAFE EPDMGMLLV DLARHAARAY S-RESNYTED EALERIVEMF ESEIARPSDL GSTTER-P-- Q
WP_1435734 MAGPK--EQP L-----PPDV VGREEAPEVL RAFVVDGGL- SIAFT-RAFE EPDMGMLLV DLARHAARAY A-RESDYSED EALARIVDMF DAEIARPTDV GNTTTRSQQG H
WP_1141885 HTEKKFEALN V-----PPDV -LEKGGVEIL RASVVDGAV- SIALR-RAFD DPFTWGLLV DLARHAARVY A-METDFSEE EALAEISAGI QAELDDPSDP GTTQAI---- N
WP_1509640 MPKA-FQPLS V-----PPDA -LEKGGVEVL RASVVDGAV- SVALR-RSFD DPFTWGLLI DLARHAARVY A-LETDLSEE EAFQIRAGL EAETDPPDG- PDSL-L---- N
PBB66311.1 HFGKLSSELS P-----PPN- AKNARAVEVL RVWAEPGAA- QQLVLTTHK EPGAWGLLA DVARHAARAY V-AEG-ISEA QALDRILMLF KAFAAPPTDA P-----S- -
WP_0560574 MADHP--ANH PNAIALDKGA QLTGESVEVA RIWITNGAGS NVLIDAGILE DPTVFGYLLA DTIRHAARAY A-GTWGLDED AALQAIVDGV GTELREQFTT ITTIQEGM-- H
WP_1504275 HFGKKGIELE E-----PPL- SRDAGAFEIL RVWGGDNL- QQYSLKTVRD DPGAWGLMLV DIARHVAKAY G-NTGDFSEE AALKRIKELF DAENASPTDT PLQV----K- -
WP_0489346 MKS--ELP-- I-----PPAA LQDVDSKELL RVWAAAGNQ- HISIATGVWE NPTINGIMLV DLLRHARSY E-KVGNVSYE ESMRLIKAGF DAEWESPTD- ----- -
WP_1334671 MAGPK--EQP L-----PPDV MGRDDAVEVL RAFVVDGGL- SIAFQ-RAFE EPDMGMLLV DLARHAARAY S-RESEYTED EALARIVEMF EAEIARPTD- ----- -
WP_1588214 MSGNK--VLD P-----PPAA VRDKASFELL RVWVAEQGQ- HVSLRPGTWD DPFAMGIVLA DLARHIVNAE SIHRKNFDED AFLERMLEGF RAEIESPTDD PEGEIM--Q- -
TGN80551.1 MAGPK--EQP L-----PPDV MTRDDAVEIL RVFVLDGGL- SMAFQ-RAFE EPDMGMLLV DLARHAARAY A-RESEYTEE DALNRILDMF QAEIERPTDT GTTTPRGKG- H
    
```

Figura S5. Alineamiento múltiple en formato Phylip obtenido en T-Coffee.

A4 - Material suplementario: Implicancias del receptor ionotrópico NMDA subunidad 3A en la esquizofrenia

HMMER



Distribution of Significant Hits

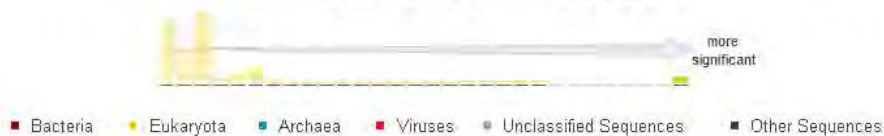


Figura S1. Dominios, estructura secundaria, y desorden predichos. En el gráfico superior se muestran las predicciones de regiones desordenadas, hélices transmembranas, péptidos señal, y dominios reconocidos. Además se puede observar en el gráfico inferior, una distribución únicamente eucariota para los homólogos de esta proteína.

Quick2D



Figura S2. Predicción de estructura secundaria. La estructura secundaria predicha para Q8TCU5 es mayormente hélice alfa combinada con hoja beta plegada, en la región inicial es más flexible que el resto de la proteína por tener un gran desorden, al igual que la región c-terminal. La región c-terminal posee además un segmento coiled coil predicho. En la región intermedia nuevamente se observa la predicción de segmentos transmembrana (cuatro segmentos transmembrana).

Nota: Estructura secundaria: hélice-alfa=color rosa, hoja plegada beta=azul, n-hélice=color rojo, coiled coils=color verde, transmembrana=color fucsia, Desorden= color beige.

PSIPRED tipos de aminoácidos, y regiones transmembranales.

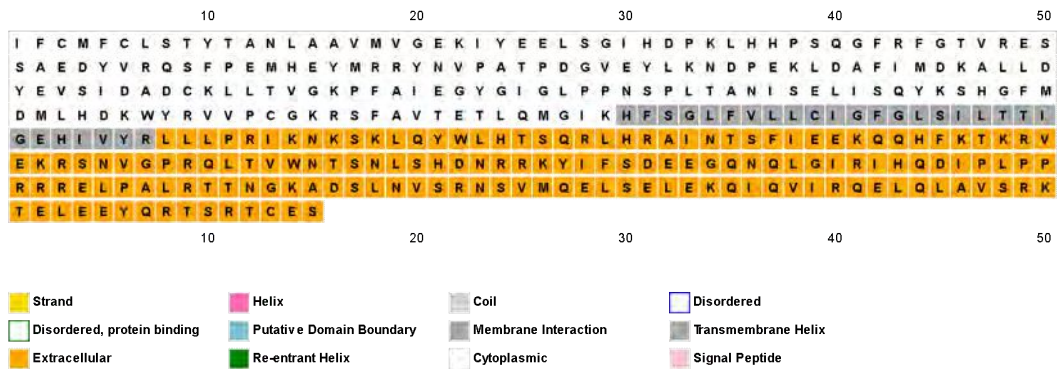


Figura S3. Topología de la región C-terminal. Se observa que la región C-terminal posee un segmento transmembranal.

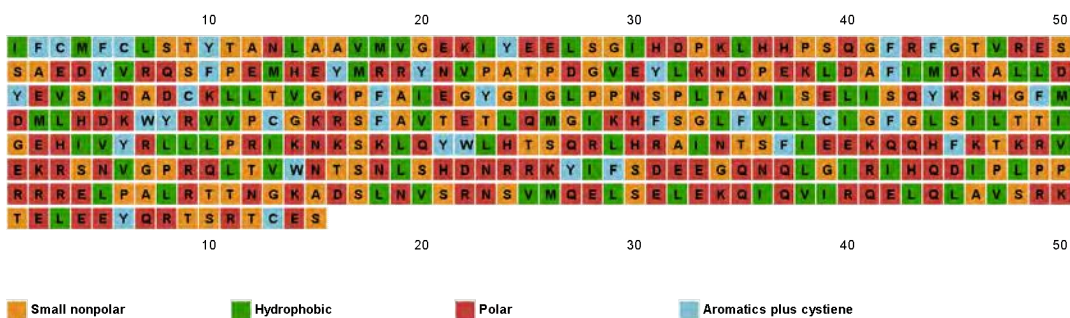


Figura S4. Polaridad de cada posición dada por PSIPRED. Los aminoácidos de la región (1000-1115) son mayormente polares, lo que no concuerda con el hecho de que la región este compuesta de hélices transmembranales y segmentos interactuantes con la membrana, ya que estos residuos se esperaba que fueran no polares, y a su vez contradice el hecho de que la región fuera predicha como globular por el servidor GlobPlot. Esto confirma el hecho de que la región posee una alta variabilidad secuencial por ser una región desordenada (tal como indican las predicciones de IUPred2A, Dynamine, y Pfam), lo que al parecer dificulta enormemente las predicciones que se hagan sobre ella.

PSIPRED dominios predichos en region c-terminal

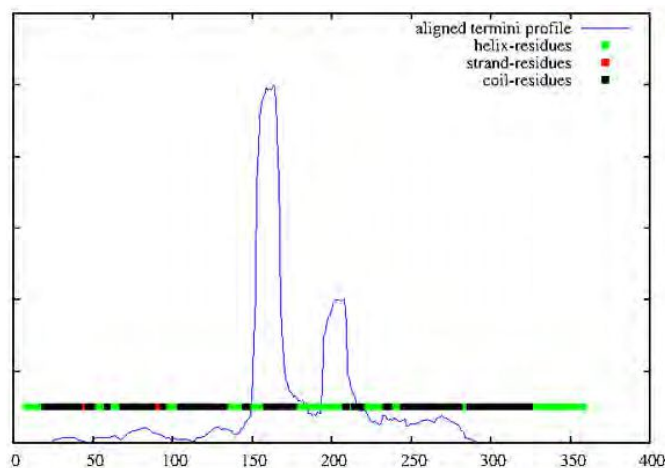


Figura S5. Dominios detectados. Se observa los dominios detectados de la región c-terminal como dos picos en la gráfica. Los dominios podrían ser propios de la interacción entre la proteína y la membrana, ya que la posición en que se ubican ambos (1050-1110 aproximadamente) fue definida como transmembranal anteriormente por servidores como PSIPRED. Límites de dominio putativos ubicados en el perfil de alineación PSI-BLAST: Número de dominios previstos por DPS= 2; Ubicaciones de límites de dominio predichas DPS= 164.

TMPRED

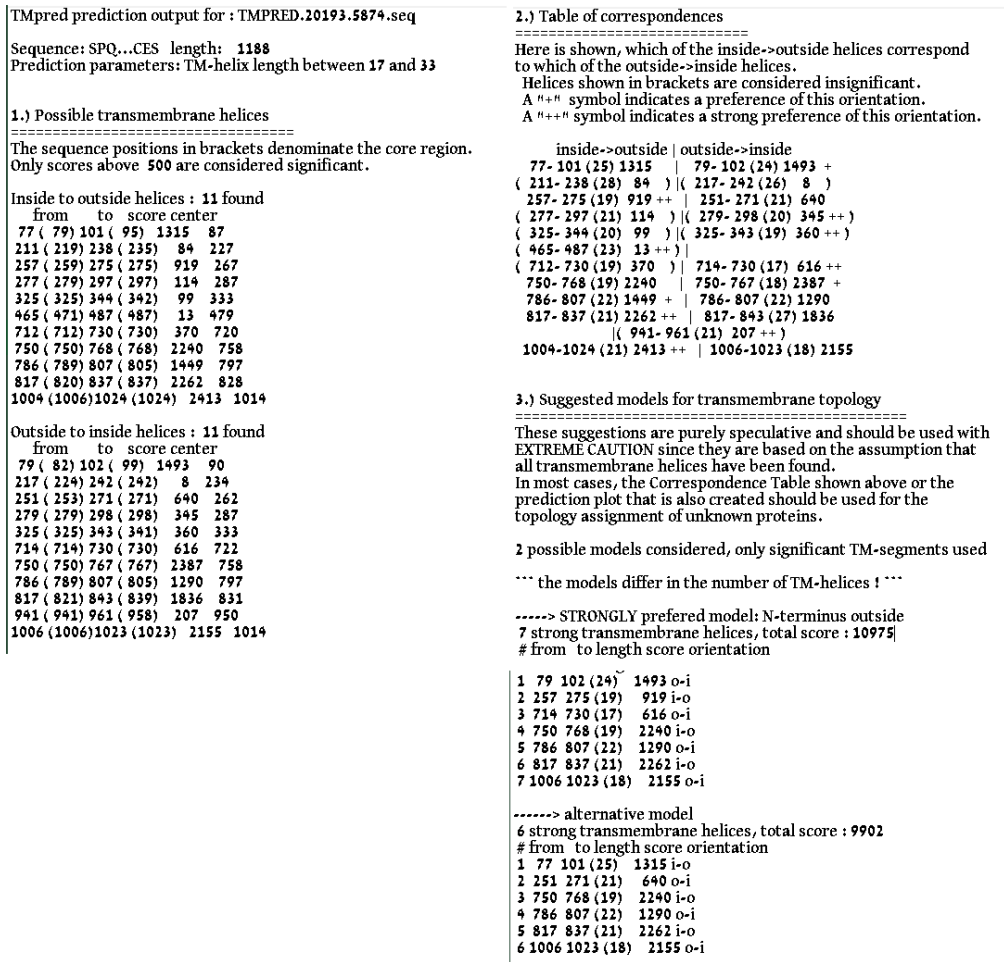


Figura S8. Resultados de la predicción de hélices citoplasmáticas y no citoplasmáticas en la secuencia proteica de Q8TCU5.

DynaMine

Protein: sp|Q8TCU5|NMD3A_HUMAN Glutamate receptor ionotropic, NMDA 3A OS=Homo sapiens OX=9606 GN=GRIN3A PE=1 SV=2

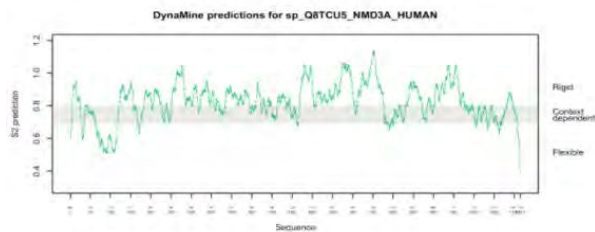


Figura S9. Predicción de desorden. La figura presenta un gráfico, a la izquierda, con las regiones rígidas y flexibles de la proteína que fueron predichas, así como también se muestra a la derecha, de manera representativa cuales regiones de la secuencia tienen mayor (rojo) o menor (azul) desorden, con una escala de colores intermedios para los valores intermedios, en base a la predicción. Nuevamente se observa una región altamente desordenada al inicio de la proteína (1-200), seguida de un segmento muy conservado de la proteína (donde se encontraban las regiones transmembrana), y por último se encuentra un segmento pequeño desordenado hacia el final de la proteína (1100-1115). El resultado obtenido es el de una proteína poco flexible, bien conservada, con poco desorden. Resultados obtenidos con parámetros por default.

MobiDB

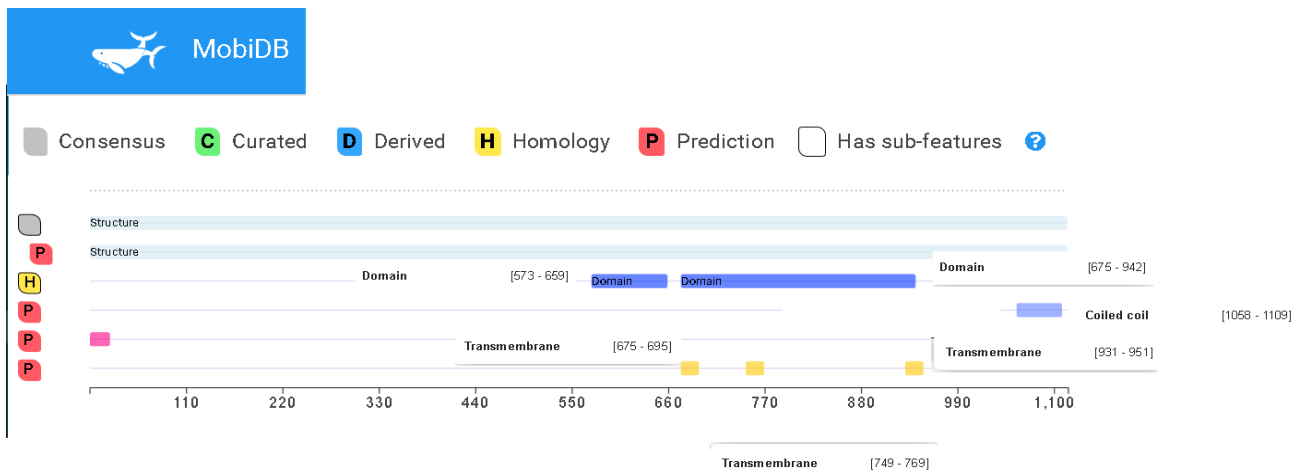


Figura S10. Resultados de MobiDB, se utilizaron los parámetros por default. Como se observa en la figura, se indica la presencia de regiones transmembrana en el segmento que va del 675 al 695, el 770 al 931, y del 931 al 951, lo que concuerda con las predicciones arrojadas por los otros servidores (Paircoil, Paircoil2, Phobius, COILS, TMHMM, TMPred). También se indica la presencia de un péptido señal (color rosa) en la región N-inicial de la secuencia.

Resultados de Pfam

Source	Domain	Start	End
sig_p	n/a	1	26
low_complexity	n/a	2	21
low_complexity	n/a	13	26
disorder	n/a	26	28
disorder	n/a	36	37
disorder	n/a	59	118
low_complexity	n/a	61	74
Pfam	ANF_receptor	124	471
low_complexity	n/a	157	181
low_complexity	n/a	548	567
Pfam	Lig_chan-Glu_bd	557	661
Pfam	Lig_chan	674	942
transmembrane	n/a	675	695
transmembrane	n/a	716	734
transmembrane	n/a	746	769
transmembrane	n/a	934	956
disorder	n/a	998	1001
disorder	n/a	1025	1026
disorder	n/a	1040	1043
disorder	n/a	1045	1046
low_complexity	n/a	1047	1059
disorder	n/a	1049	1074
coiled_coil	n/a	1065	1099
disorder	n/a	1100	1101
disorder	n/a	1104	1115

Figura S11. Dominios detectados con Pfam. Se detectaron 3 familias analizando la secuencia proteica de Q8TCU5; ANF receptor, Lig chan-Glu bd, y Lig chan.

Resultados de CDART.

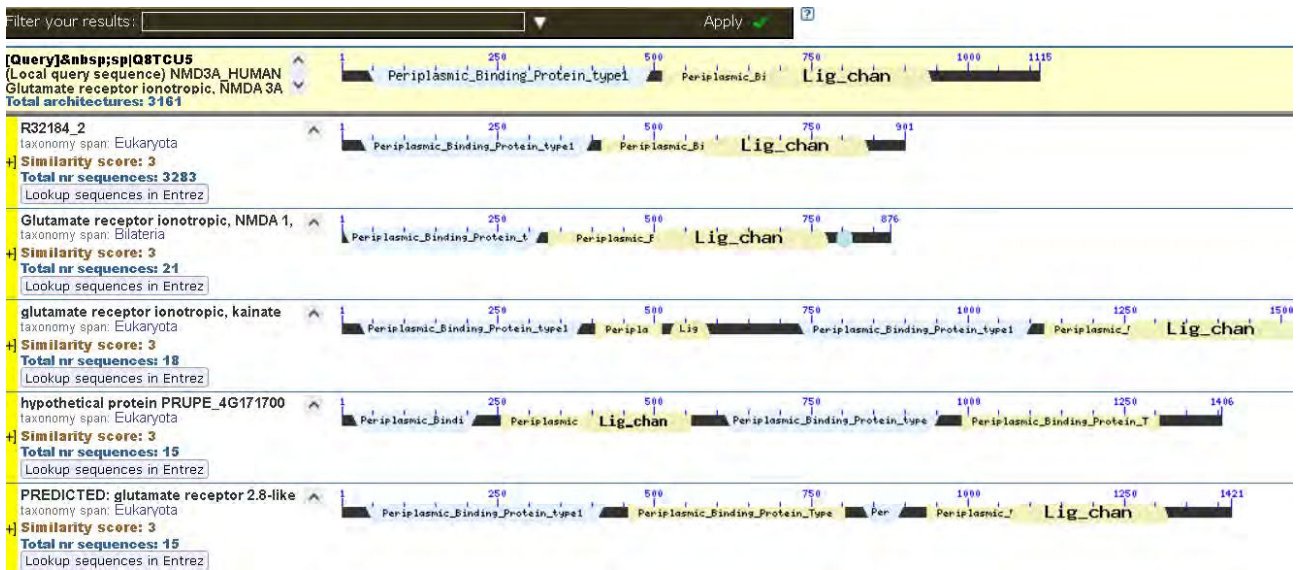


Figura S12. Dominios detectados por la base CDART. Arquitecturas coincidentes encontradas para la proteína Q8TCU5, con sus respectivas taxonomías, y secuencias con dominios similares.

PHYRE2



Figura S13. Modelo final. Imagen coloreada N → C terminal. Dimensiones del modelo (Å): X:284.116 Y:234.269 Z:331.922. 74% de los residuos modelados con un > 90% de confianza.

Pymol RMSD de los templates elegidos

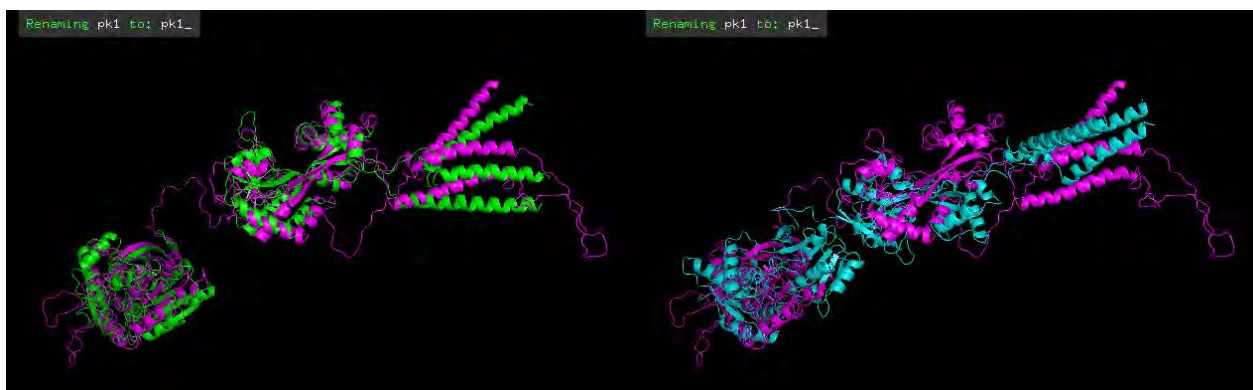


Figura S14. (a) A la derecha se exhibe el alineamiento estructural obtenido entre el modelo obtenido (UKNP.B99990006) en color fucsia, y el template 7KS0_D en color verde, cuyo RMSD fue de 4.103. (b) En la imagen izquierda se encuentra el alineamiento estructural entre el modelo obtenido (UKNP.B99990006) en color celeste, y el template 7KS0_C, en color fucsia, con un RMSD de 21.494.

Polaridad del modelo UKNP.B99990006

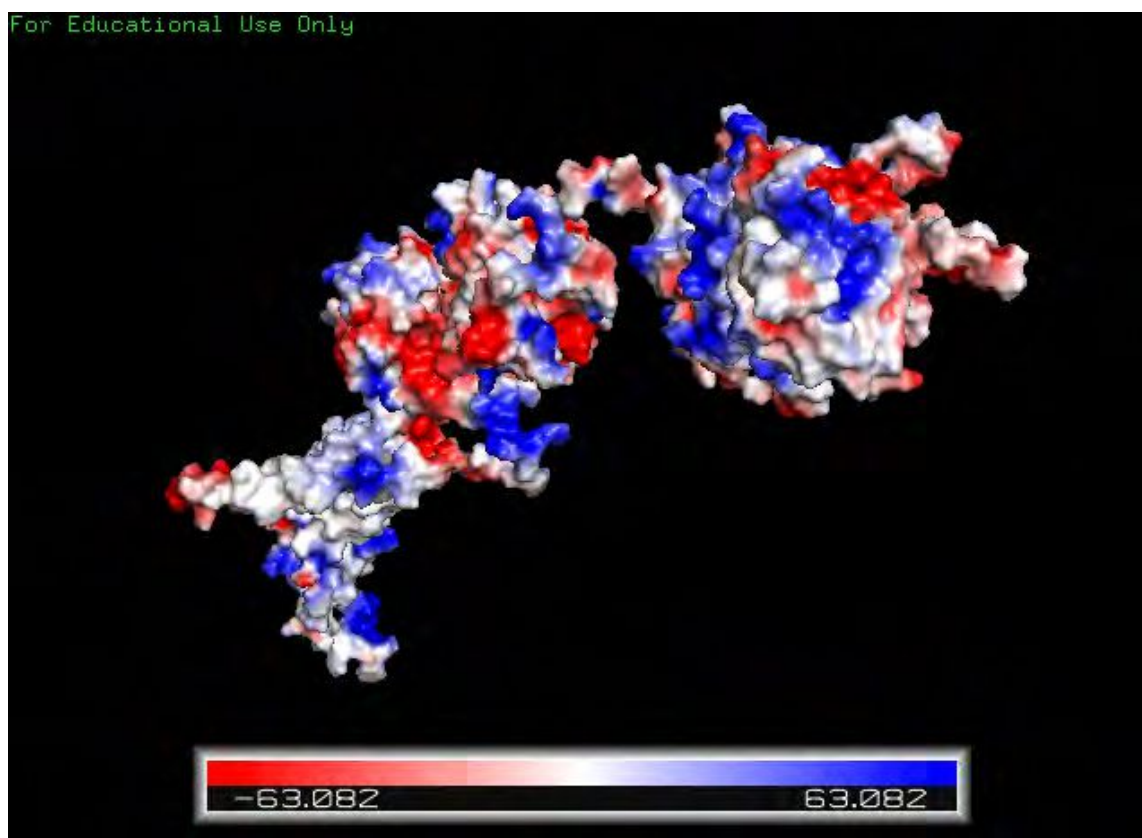


Figura S15. Superficie de la proteína Q8TCU5 visualizada con Pymol. Superficie de la proteína modelada (UKNP.BBBB99990006), se muestra la valencia de la proteína coloreada en color rojo para las cargas negativas, color azul para las positivas, y blanco para los segmentos neutros.

T-COFFEE bloques que presentan mayor variabilidad

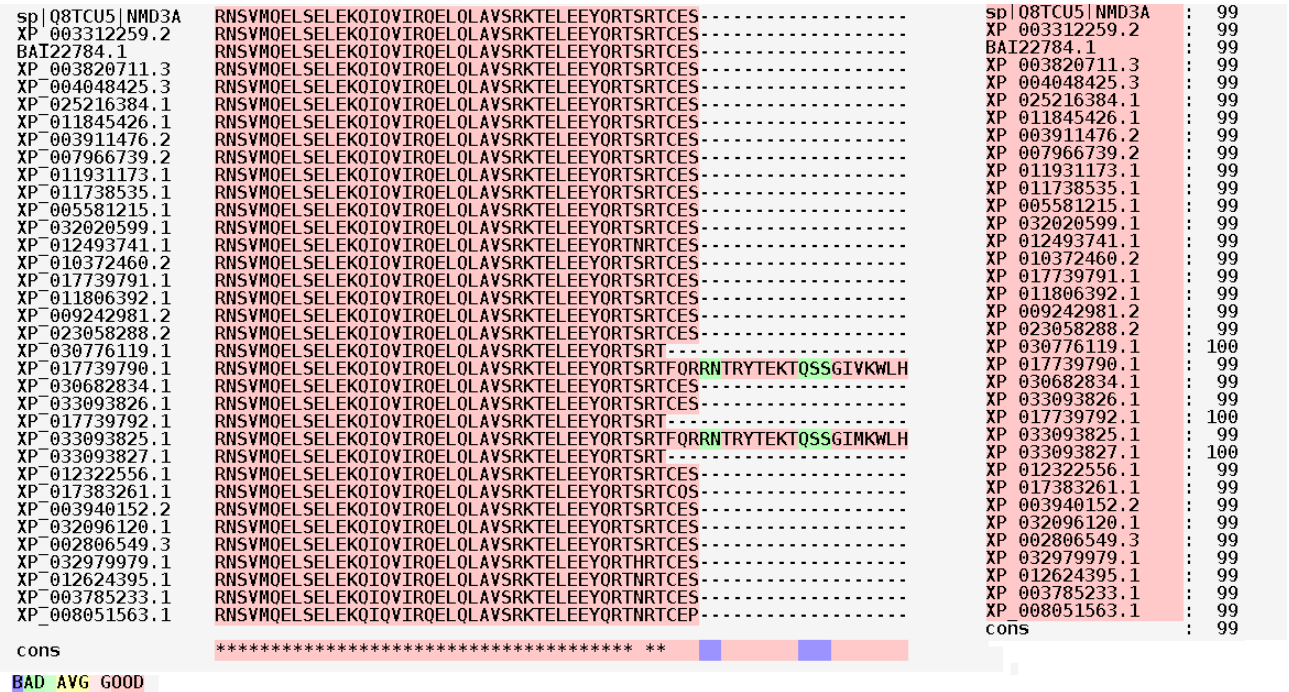


Figura S16. En la figura se muestran las regiones C-terminales de las secuencias de primates, se puede observar dos regiones C-terminales distintas para los organismos *Trachy pithecus*, y *Rhinopithecus bieti*.

CONSURF

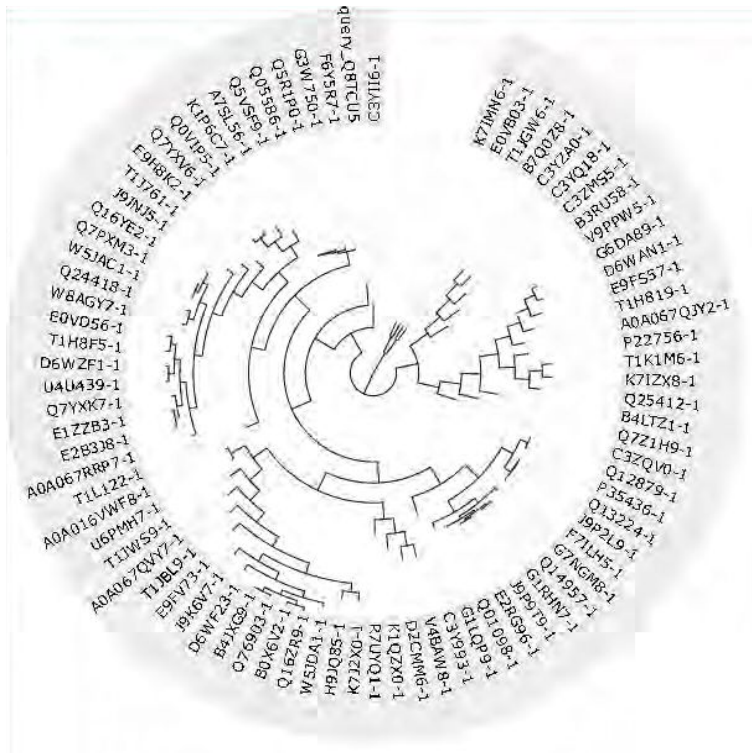


Figura S17. Árbol obtenido en Consurf por el método de máxima similitud.

Consurf: variabilidad de las secuencias

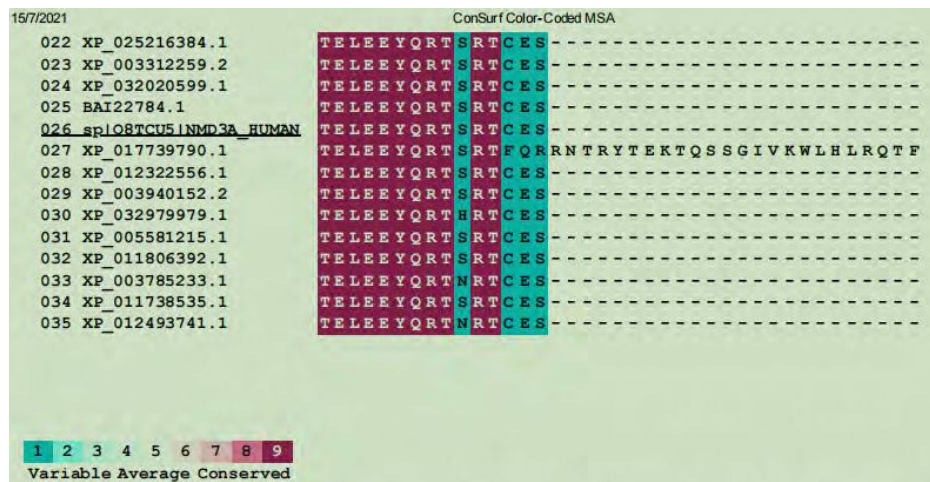


Figura S18. En la figura se resaltan las regiones conservadas (color violeta) y no conservadas (color celeste) de las secuencias de Primates. Como se puede observar, las regiones terminales son altamente variables, y este segmento en *H. sapiens* parece poseer una longitud mayor en comparación con el resto de los Primates.

Segmento c-terminal dominios intra y extracelulares

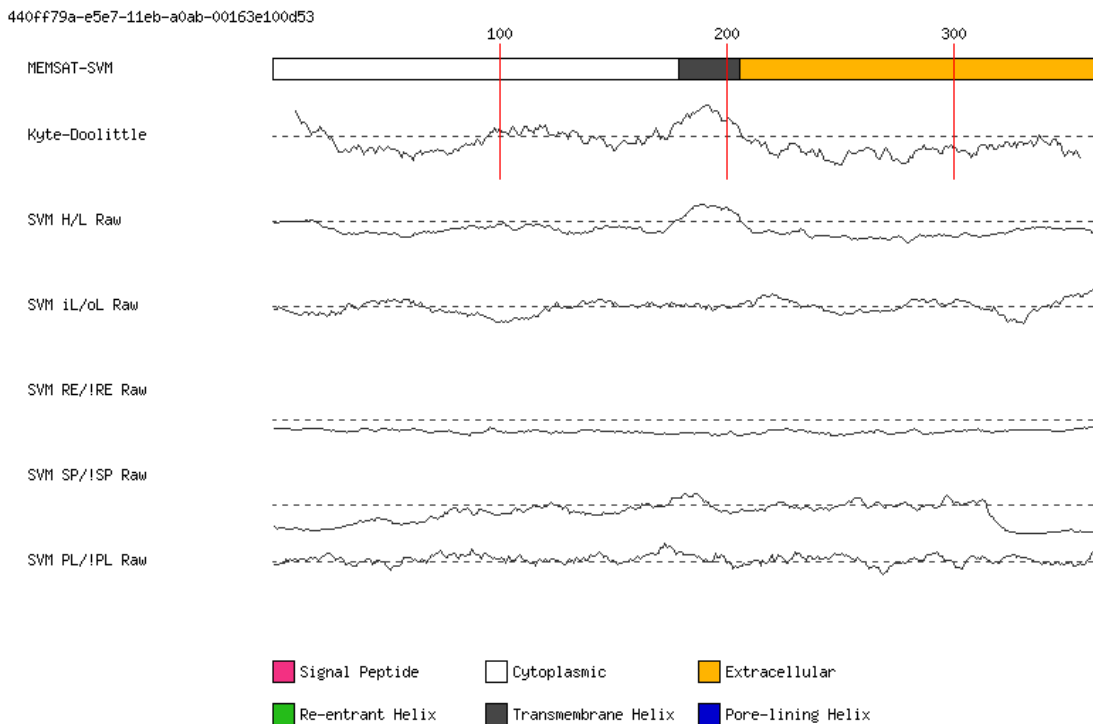


Figura S19. En la figura se exhiben las predicciones para el segmento c-terminal de Q8TCU5, se predice que el mismo es extracelular.

A5 - Material suplementario: Recorrido por Hv1: un canal selectivo de protones

En esta sección se anexan figuras de los distintos procedimientos.

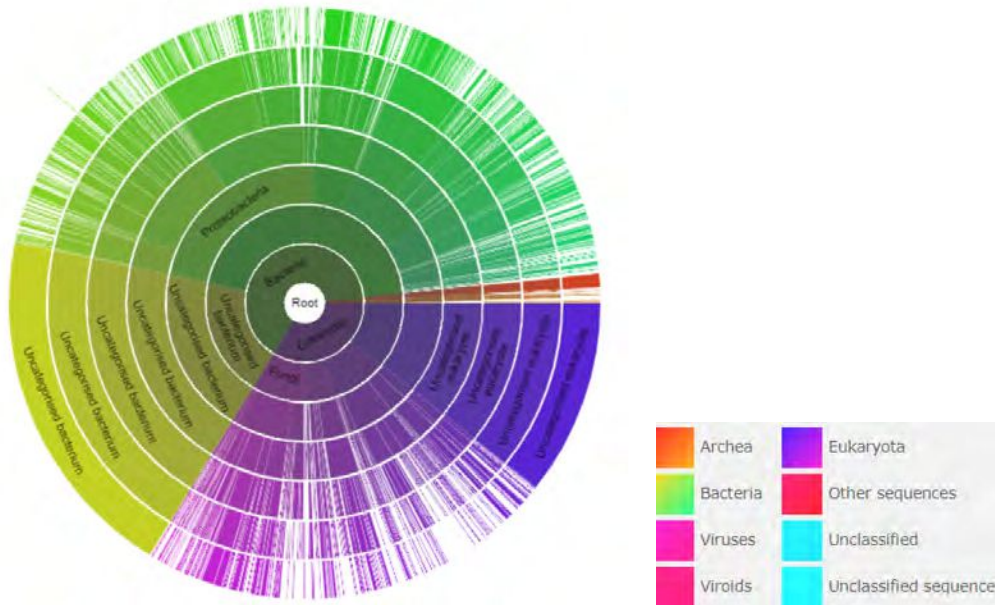


FIGURA A1. Distribución del dominio de transporte de iones en distintas especies.

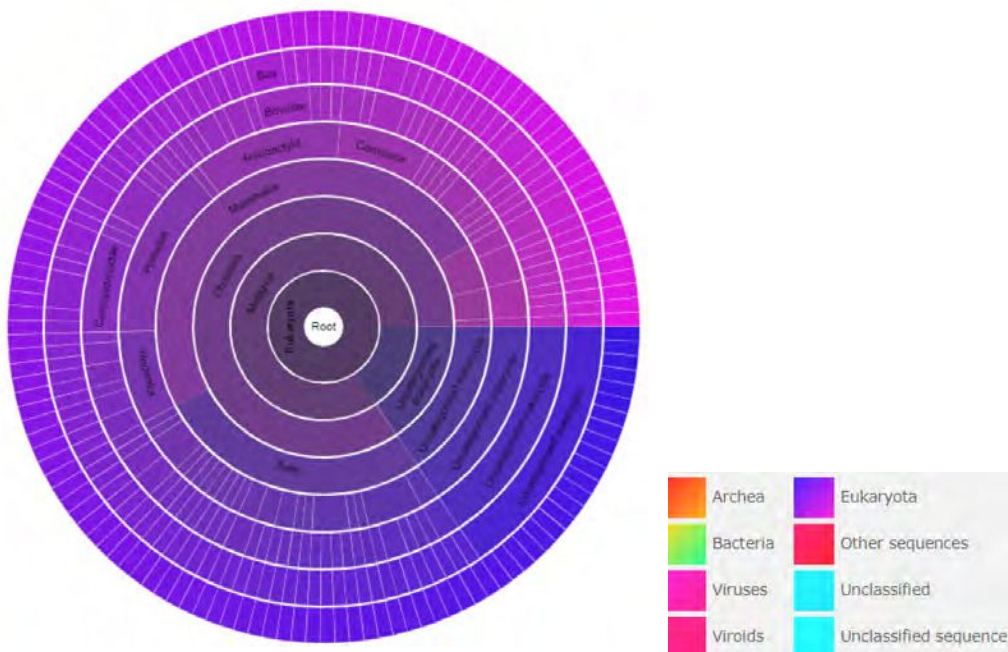


FIGURA A2. Distribución de la región C-terminal del canal de protones voltaje operado en distintas especies.

```

# sp|Q96D96|HVCN1_HUMAN Length: 273
# sp|Q96D96|HVCN1_HUMAN Number of predicted TMHs: 3
# sp|Q96D96|HVCN1_HUMAN Exp number of AAs in TMHs: 70.99847
# sp|Q96D96|HVCN1_HUMAN Exp number, first 60 AAs: 0.00035
# sp|Q96D96|HVCN1_HUMAN Total prob of N-in: 0.89918
sp|Q96D96|HVCN1_HUMAN TMHMM2.0 inside 1 100
sp|Q96D96|HVCN1_HUMAN TMHMM2.0 TMhelix 101 123
sp|Q96D96|HVCN1_HUMAN TMHMM2.0 outside 124 137
sp|Q96D96|HVCN1_HUMAN TMHMM2.0 TMhelix 138 160
sp|Q96D96|HVCN1_HUMAN TMHMM2.0 inside 161 171
sp|Q96D96|HVCN1_HUMAN TMHMM2.0 TMhelix 172 191
sp|Q96D96|HVCN1_HUMAN TMHMM2.0 outside 192 273

```

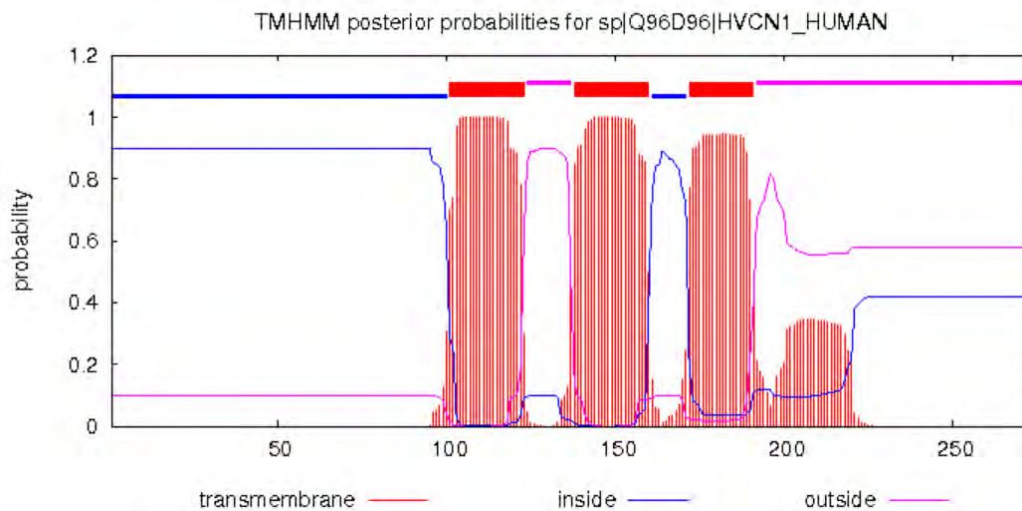


FIGURA A3. Predicción de regiones transmembrana con www.cbs.dtu.dk

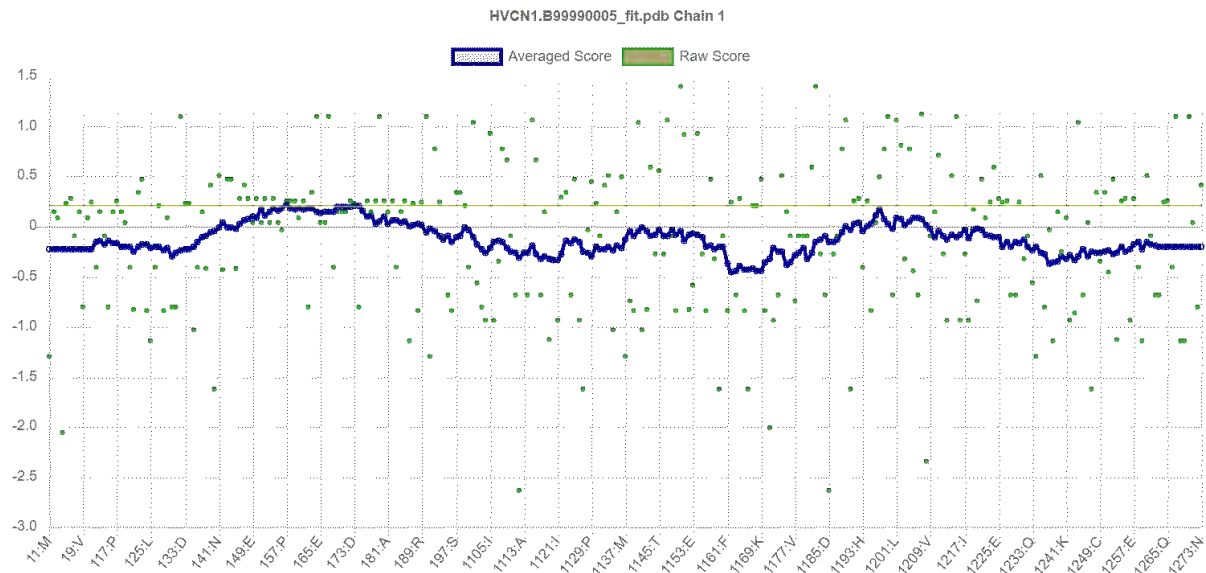


FIGURA A4. Resultado de Verify como un Plot.

A6 - Material suplementario: Análisis bioinformático de la enzima Timidilato Sintasa de Candidatus Poseidoniales archaeon

Secuencia	%Identidad	Organismo	Código NCBI
1	100	Candidatus Poseidoniales archaeon	DAC32365.1
2	88	Acidimicrobiaceae bacterium	MAP98708.1
3	91,29	Euryarchaeota archaeon	MAE78661.1
4	76	Alphaproteobacteria bacterium	NVJ71056.1
5	70	Pseudomonas aeruginosa	VFT44992.1
6	61	Alloprevotella tannerae ATCC 51259	EEX70573.1
7	55	Aureimonas sp. AU40	WP_06211400 8.1
8	50	Planctomycetes bacterium	MBK8975477.1
9	54	Candidatus Peribacteria bacterium	MSR87052.1
10	45	Delftia phage RG-2014	YP_009351922 .1
11	40	Lasallia pustulata	SLM40774.1
12	35	Candidatus Pacearchaeota archaeon	MBI2045214.1
13	30	Lentisphaerae bacterium ADurb.Bin242	OQA84413.1
14	20	Rhodobacteraceae bacterium	MBC8408728.1
15	14	deoxyuridylate hydroxymethyltransferase [Acinetobacter phage SH-Ab 15599]	AXF41386.1

Tabla S1. Características de las secuencias elegidas para el alineamiento múltiple.



Figura 1. Alineamiento de las 15 secuencias seleccionadas al azar. El recuadro en rojo marca la secuencia de la proteína de interés.

ISBN 978-950-34-2180-2



9 789503 421802