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 oligossacharides, 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 jackhmmer contra la base 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 jackhmmer 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 TCoffee.

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 (máximum 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 fosfotidilglicerol-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 <=>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.

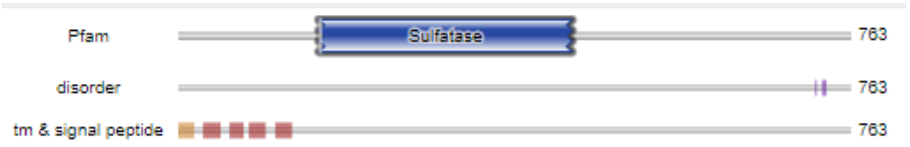

Figura 1. Características estructurales (Pfam) de la proteína mdoB de *S. agona* (Uniprot ID: B5F500).

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 fosfotidilglicerol-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 fosfatasas 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).

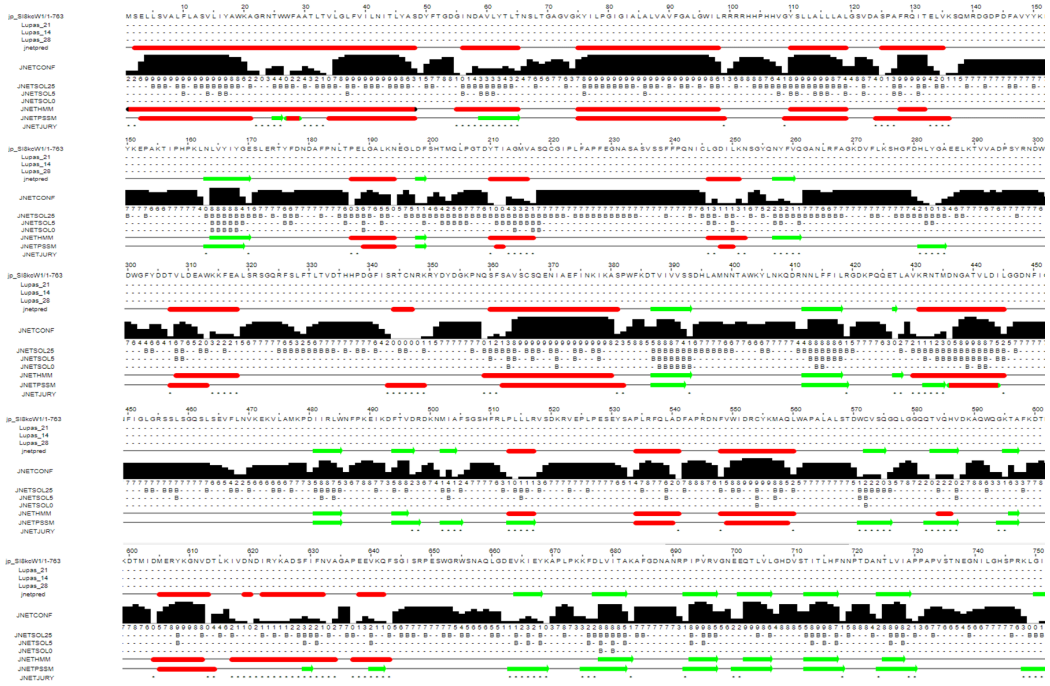

Figura 2. Resultado de la predicción de la estructura secundaria con el programa JPred. Los segmentos rojos corresponden a arreglos alfa hélices y las flechas verdes corresponden a arreglos hoja beta.

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).

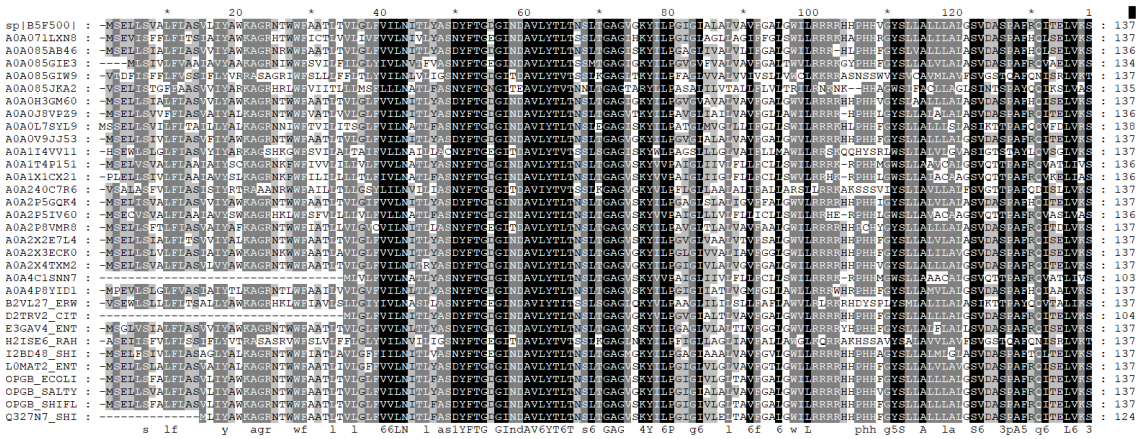

Figura 3.A. Alineamiento múltiple. La imagen corresponde a la región transmembrana conservada (59-74). Las referencias de los organismos elegidos se enlistan en el Anexo.

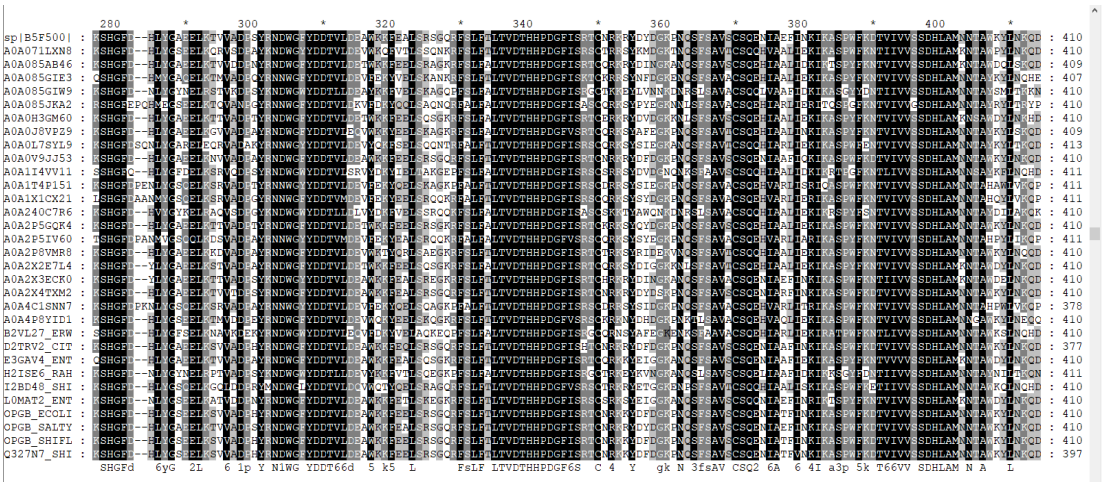

Figura 3.B. Alineamiento múltiple. La imagen corresponde al dominio Sulfatasa conservado. Las referencias de los organismos elegidos se enlistan en el Anexo.

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).

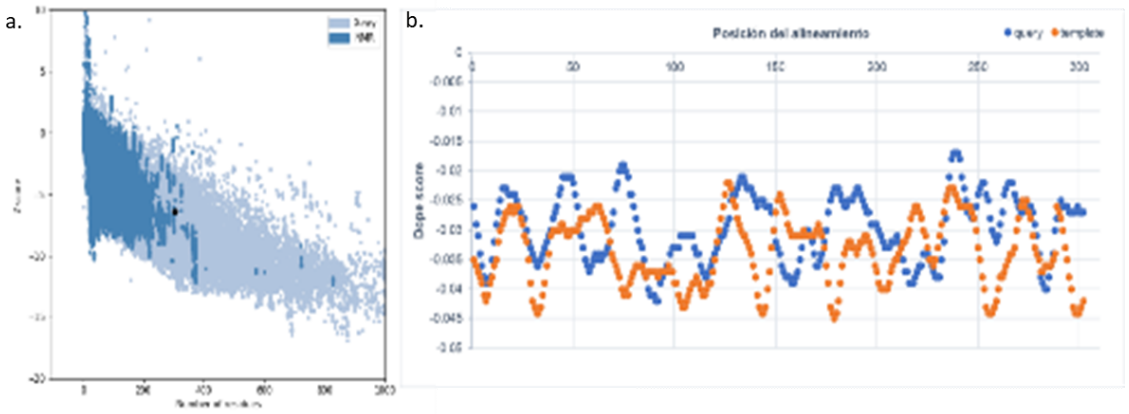

Figura 4. Análisis energético del modelo tridimensional propuesto para la proteína en estudio. a. Validación energética global por el programa ProSA. El punto negro refiere a la proteína de interés. b. Perfil energético por sitio superpuesto de las secuencias query (azul) y template (naranja).

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 RSMD arrojó un valor de 0.137, según el programa PyMol.

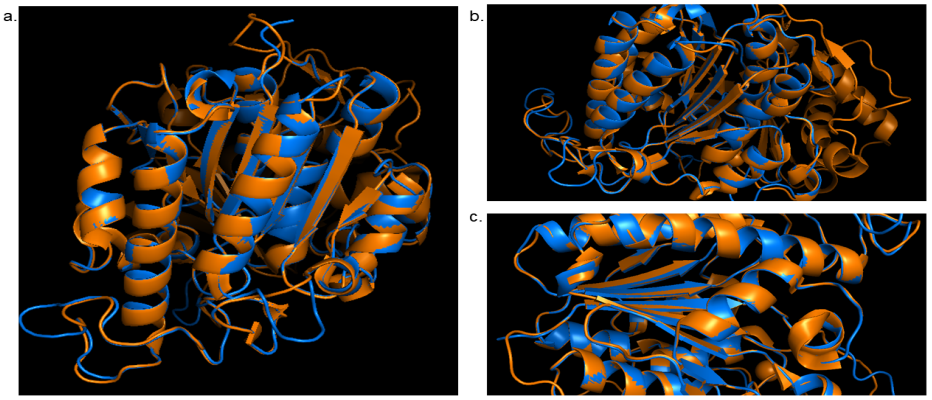

Figura 5. Alineamiento estructural. Estructura tridimensional de proteína query (azul) superpuesta con la estructura de la proteína molde (naranja). a. Se observa el plegamiento alfa hélice hacia la superficie de la estructura. En b. y c. se muestran los plegamientos hojas beta en el interior de la cavidad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

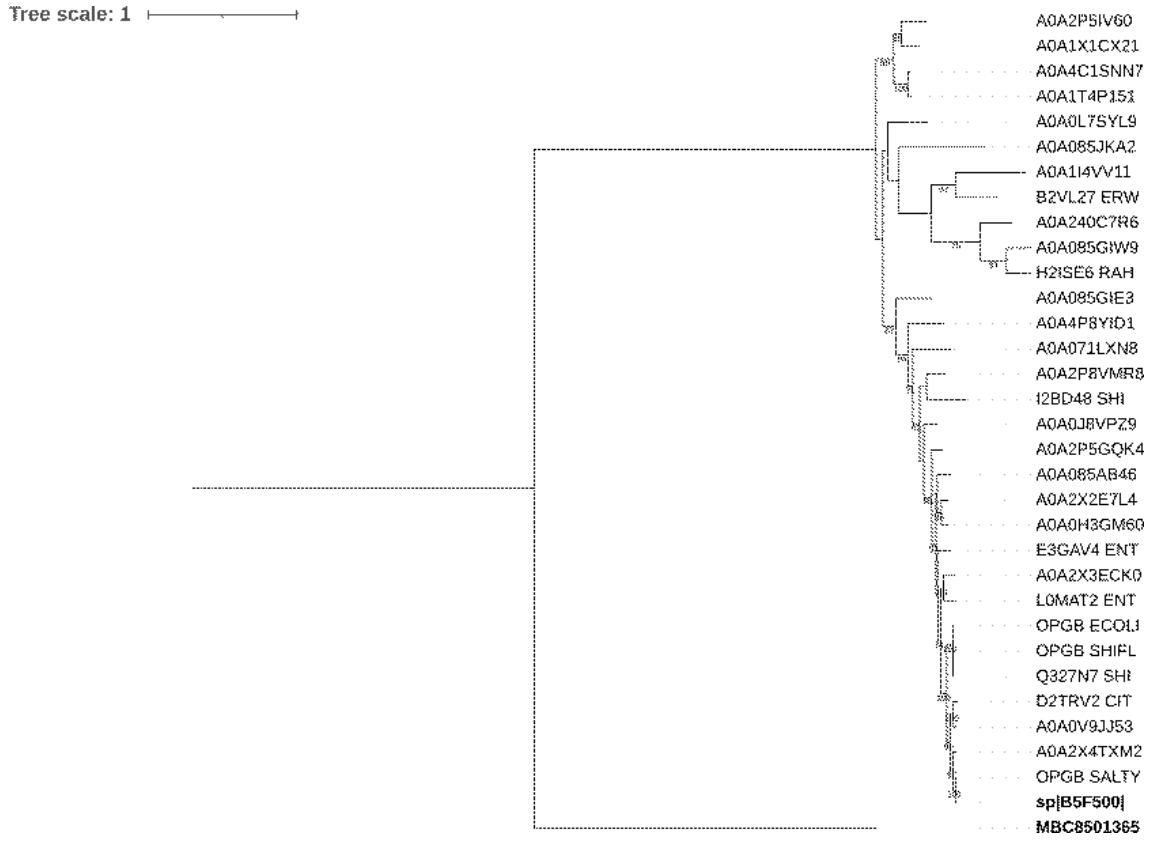

Figura 6: Árbol filogenético construido por el método de máxima verosimilitud. Bootstrap=100.Outgroup: secuencia aminoacídica de Archaea (MBC8501365). Bootstraps mostrados corresponden a valores > 50.

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.