PREDICCIÓN *IN-SILICO* DE LA ESTRUCTURA Y FUNCIÓN DE LA PROTEÍNA HIPOTÉTICA p284 DE *Trypanosoma cruzi*

MAYRA JAIMES MELO

PONTIFICIA UNIVERSIDAD JAVERIANA

FACULTAD DE CIENCIAS BÁSICAS

CARRERA DE BIOLOGÍA

BOGOTÁ D.C

2012

PREDICCIÓN *IN-SILICO* DE LA ESTRUCTURA Y FUNCIÓN DE LA PROTEÍNA HIPOTÉTICA p284 DE *Trypanosoma cruzi*

MAYRA JAIMES MELO

Ingrid Schüler García *PhD*

DECANA ACADÉMICA

Andrea Patricia Forero *MSc*

DIRECTORA DE PREGRADO

PREDICCIÓN *IN-SILICO* DE LA ESTRUCTURA Y FUNCIÓN DE LA PROTEÍNA HIPOTÉTICA p284 DE *TRYPANOSOMA CRUZI*

MAYRA JAIMES MELO

Orlando Acevedo

DIRECTOR DE TESIS

Janneth Gonzalez Santos

JURADO

NOTA DE ADVERTENCIA

"La Universidad no se hace responsable por los conceptos emitidos por sus alumnos en sus trabajos de tesis. Solo velará por que no se publique nada contrario al dogma y a la moral católica y por que las tesis no contengan ataques personales contra persona alguna, antes bien se vea en ellas el anhelo de buscar la verdad y la justicia".

Artículo 23 de la Resolución No 13 de Julio de 1946.

AGRADECIMIENTOS

A Orlando Acevedo por haber aceptado ser mi director de grado y por toda su ayuda y aportes.

A Janneth Gonzales quien despertó en mí el interés por la Bioinformática.

A mi familia, porque sin su apoyo no habría logrado realizar este trabajo de grado.

Al Laboratorio de Bioquímica molecular computacional y Bioinformática de la Pontificia Universidad Javeriana por todas sus enseñanzas.

TABLA DE CONTENIDO

1. RESUMEN

La enfermedad causada por el parásito *Trypanosoma cruzi* denominada *mal de Chagas* constituye un problema de salud de grandes proporciones en Latinoamérica. A pesar de que se han realizado numerosos estudios moleculares sobre este parásito aún se desconocen las funciones de algunas de sus proteínas y esta información puede ser clave para el control, diagnóstico y prevención esta enfermedad. Una parte importante de las proteínas desconocidas de *T.cruzi* está constituida por *proteínas hipotéticas*, es decir, proteínas de las cuales sólo se conoce su secuencia de aminoácidos, pero cuya existencia no ha sido comprobada experimentalmente.

Con el fin de contribuir a complementar la información existente sobre el proteoma de este parásito se realizó la predicción *In-silico* (por medio del uso de computadoras) de la estructura y función de la *proteína hipotética p284* de *T. cruzi* utilizando varios tipos de software bioinformático. En primer lugar se comparó la secuencia de la proteína obtenida de NCBI (National Center for Biotechnology Information) con las secuencias de las proteínas consignadas en las bases de datos utilizando el programa BLAST (Basic Local Alignment Search Tool); posteriormente se determinaron sus características fisicoquímicas por medio del programa *ProtParam*, las características de su estructura secundaria por medio de GORIV, el tipo de plegamientos de su estructura terciaria por medio de SCOP (Structural Classification Of Proteins), y sus regiones conservadas por medio de *Motif Search*. También se obtuvieron varios modelos de la estructura tridimensional de la proteína utilizando plantillas de proteínas similares obtenidas de PDB (Protein Data Bank) por medio del uso de programas como SAS (Sequence Annotated by Structure), I-TASSER y (PS)2.

Los resultados indican que la *proteína hipotética p284* pertenece al grupo de las proteínas *MutS2* implicadas en eventos de recombinación, reparación y dimerización del ADN. También se encontró que la posible estructura tridimensional de la proteína es muy similar a la de otras proteínas que contienen el dominio *Smr (small mutations related domain)* que es característico de las proteínas *MutS2*.

2. INTRODUCCIÓN

La secuenciación del genoma de *Trypanosoma cruzi,* el agente causal del mal de Chagas, fue y sigue siendo una herramienta de sumo valor para el estudio de diversos aspectos de la enfermedad pues puso a disposición de los investigadores de este parásito una gran cantidad de información acerca de sus genes, y por consiguiente de sus proteínas (1,2). El estudio de la composición protéica de este organismo es indispensable para comprender sus mecanismos infectivos y patogénicos, pues muchos

de ellos implican la interacción de diversas proteínas y enzimas (1,2,3). Una parte importante de las estructuras proteínicas de *T.cruzi* está constituida por *proteínas hipotéticas*, denominadas así porque la única información disponible sobre ellas es la de su secuencia de aminoácidos que es obtenida a partir de la traducción directa de posibles regiones codificantes de un genoma secuenciado (4). Durante los años 60 se desarrollaron diferentes técnicas bioinformáticas para predecir la estructura y función de estas secuencias generando una revolución en la forma en que se estudian las proteínas, pues por medio del uso de estas herramientas es posible caracterizarlas sin necesidad de acudir a complicados procedimientos experimentales (4,5).

En este estudio se utilizaron diferentes programas de análisis bioinformático con el objetivo de predecir *in silico* la función y estructura de la *proteína hipotética p284* de *Trypanosoma cruzi.* Los resultados obtenidos pueden ser de gran importancia pues indican que la función de esta proteína está relacionada con la reparación y recombinación del ADN de este parásito.

3. PLANTEAMIENTO DEL PROBLEMA Y JUSTIFICACIÓN

La presencia del parásito causante del *mal de Chagas*, *Trypanosoma cruzi*, en las zonas tropicales de Latinoamérica genera un gran impacto en la salud de la población humana que las habita, pues millones de personas se ven afectadas por esta enfermedad (6). La importancia de la parasitosis radica en su elevada ocurrencia y en las grandes pérdidas económicas generadas por incapacidad laboral y muerte repentina de personas aparentemente sanas (2,6). En 1993 el Banco Mundial ubicó a la enfermedad de Chagas en el primer lugar de prevalencia entre las enfermedades tropicales y en el cuarto entre las enfermedades transmisibles en Latinoamérica, sólo debajo de las infecciones respiratorias agudas, de las enfermedades diarreicas y del SIDA (2). A pesar de que desde entonces han aumentado los controles para prevenir el contagio de la población, y de que hay numerosos estudios sobre la biología de *T.cruzi*, esta enfermedad sigue siendo un problema grave para los países centro y suramericanos (2,6). Debido a la gran cantidad de inconvenientes que genera este parásito es necesario complementar la información existente acerca de sus características moleculares, pues muchos de estos datos se pueden utilizar para mejorar el diagnóstico, tratamiento y prevención de la enfermedad (1,2,6). Una parte importante de la investigación de los aspectos moleculares de este organismo consiste en caracterizar sus *proteínas hipotéticas* de las cuales sólo se conoce la secuencia de su estructura primaria, es decir, su secuencia de aminoácidos (5). El objetivo de este trabajo es predecir por medio de diferentes herramientas bioinformáticas la estructura y función de la *proteína*

hipotética p284 de *T.cruzi* con el fin de complementar la información existente sobre el proteoma de este importante parásito.

4. MARCO CONCEPTUAL

Trypanosoma cruzi es un protozoo perteneciente al grupo de los *kinetoplástidos,* caracterizados por poseer ciclos de vida muy complejos y por ser agentes de graves enfermedades en las zonas tropicales de Centro y sur América (6,7). *T. cruzi* es el causante de una de las enfermedades más graves y extendidas en el continente americano, la *enfermedad de Chagas*, denominada así en honor a su descubridor *Carlos Chagas* (2,6,7). Esta enfermedad se transmite a los mamíferos (incluyendo al hombre) a través del contacto con heces infectadas de insectos Triatominos (Hemípteros), y sus consecuencias pueden ser fatales (2,7). Se calcula que esta enfermedad afecta a más de 15 millones de personas y que más del 25% de la población de Latinoamérica está en riesgo de contagio (2,7). Este grave problema de salud genera grandes pérdidas económicas pues miles de personas se ven incapacitadas para trabajar gracias a las complicaciones cardiacas y digestivas y a los daños mentales causados por esta enfermedad (2,6,7). Debido a la gran cantidad de problemas que genera la parasitosis de *T. cruzi* se han realizado numerosos estudios moleculares con el fin de mejorar el diagnóstico, tratamiento y prevención de esta enfermedad, sin embargo, aún quedan muchos aspectos por investigar. Uno de los aspectos que constituye un vacío teórico en el conocimiento de el proteoma *de T.cruzi* son sus *proteínas hipotéticas,* es decir, proteínas que han sido predichas a partir de la secuencia de los genes que las codifican, pero para las cuales no existe evidencia experimental (5); la única información que existe sobre ellas es la de las secuencias de aminoácidos que las conforman (4,5).

Hace varios años los únicos procedimientos que permitían determinar la estructura y la función de una proteína eran *la secuenciación, la cristalografía de rayos X* y *la resonancia magnética nuclear*, pero gracias a los avances en bioinformática durante los años 60, se desarrollaron diferentes tipos de software que permiten hacer comparaciones entre secuencias de aminoácidos que se encuentran en bases de datos como *genebank (banco de genes) y NCBI* (National Center for Biotechnology Information), para así determinar la estructura y función de proteínas sin tener que acudir a un laboratorio (8,9). El desarrollo de estos nuevos mecanismos de predicción ha permitido amortiguar el gran problema teórico que constituye el desconocimiento de la función y estructura de miles de *proteínas hipotéticas* (9). Cuando se realizan comparaciones de una secuencia de una *proteína hipotética* con otras proteínas con la ayuda de estos programas bioinformáticos, es posible inferir una función si los resultados muestran similitudes significativas en regiones específicas altamente

conservadas de las proteínas, pues es posible que una secuencia con función desconocida sea similar a otra secuencia cuya función ya haya sido determinada (9,10)

Es importante saber que por medio del análisis de las secuencias de aminoácidos se pueden detectar dos tipos de estructuras fundamentales para determinar la función de una proteína; los *motivos y* los *dominios. Los motivos* son secuencias de *aminoácidos* conservadas (es decir que son muy similares en proteínas relacionadas) que se pliegan de formas específicas y que sirven para dar estabilidad y funcionalidad a las proteínas (8,10,11). *Los dominios* también son secuencias conservadas que se pliegan independientemente del resto de la proteína y que constituyen regiones diferenciadas dentro de su estructura; de hecho es la asociación de los distintos dominios la que origina la *estructura terciaria* (11). La detección de similitudes en los *motivos y dominios* de las proteínas puede indicar la presencia de *homologías*, es decir, de un origen evolutivo común (11). Las investigaciones sobre este tema aseguran que la inferencia de una homología es la conclusión más poderosa que se puede establecer para indicar similitudes, ya que las proteínas homólogas comparten estructuras tridimensionales similares (10,11).

Hoy en día el uso de software bioinformático para el análisis de secuencias está extendido por todo el mundo en gran parte gracias a que en internet es posible conseguir gran variedad de programas de uso gratuito que permiten analizar secuencias protéicas que se encuentran de bases de datos de acceso libre (5,10). Muchos de estos programas son creados por centros de investigación de gran prestigio, o por empresas farmacéuticas interesadas en investigar las interacciones de las proteínas con diferentes componentes químicos (8,10). Estos programas de predicción se basan en *algoritmos matemáticos* que permiten encontrar la solución a un problema entre miles de soluciones de una forma eficiente (5,9). Entre los programas más utilizados se encuentran los siguientes:

- *BLAST* (Basic Local Alignment Search Tool): Es una herramienta que realiza alineamientos de secuencias para encontrar similitudes entre ellas. Dentro de este software se pueden escoger diferentes algoritmos para analizar las secuencias con diferentes grados de sensibilidad (BLASTp, PSI-BLAST, etc) (12,13). El programa es capaz de comparar una secuencia problema (también denominada en la literatura secuencia *query*) contra una gran cantidad de secuencias que se encuentren en una base de datos. Si se encuentran similitudes entre los *motivos y dominios* de la secuencia *query* y los de alguna de las proteínas ya conocidas se puede determinar la existencia de homologías, es decir, de un origen evolutivo común entre ellas (14,15)

- *Protparam* (Características físicas y químicas): Es una herramienta de *ExPASy* (un servidor del Instituto de Bioinformática de Suiza) que predice las características físcoquímicas de una proteína

como el número y la proporción de aminoácidos, la hidropaticidad (ver si la proteína es hidrofóbica o hidrofílica), el punto isoeléctrico (pH al que la proteína tiene carga neta de cero) y la estabilidad en un tubo de ensayo (16)

- *GOR* (Secondary Structure Prediction): es un programa que utiliza algoritmos para predecir el porcentaje de láminas Beta y hélices Alfa que constituyen la estructura secundaria de la proteína (17)
- *Motif Search*: es un programa que utiliza algoritmos para predecir si existen similitudes entre los motivos de secuencias de la proteína problema y los motivos de proteínas en las bases de datos (18)
- *SCOP* (Structural Classification of Proteins): este programa permite predecir cuáles son los motivos estructurales de la proteína, es decir, predice cómo se plegarían estas cadenas de aminoácidos conservadas (19).
- *SAS* (Sequence Annotated by Structure), $(PS)^2$ (Protein Structure Prediction Server), I-TASSER: son herramientas para aplicar información estructural a una secuencia proteínica dada. Utilizan la secuencia FASTA de una proteína dada y la compara con todas las proteínas de estructura 3D conocidas en *Protein Data Bank* (PDB) (20,21,22). Los alineamientos múltiples resultantes pueden ser coloreados de acuerdo a sus características estructurales y las estructuras 3D que concuerden se pueden superponer y visualizar en RasMol y Swiss PDB viewer (22).
- *RasMol y* Swiss PDB viewer: software de visualización Molecular.

En Colombia existen actualmente varios centros de investigación en bioinformática en importantes Universidades como la Pontificia Universidad Javeriana, la Universidad de los Andes, la Universidad Nacional de Colombia, entre otras. También se realizan estudios de este tipo en instituciones científicas como Cenicafé y FIDIC (Fundación Instituto de inmunología de Colombia) (9). A pesar de esto es necesario fomentar y extender el uso de la bioinformática en las diferentes instituciones y centros educativos de carácter científico pues puede ser aplicada en innumerables investigaciones de tipo molecular (9).

5. PREGUNTA DE INVESTIGACIÓN

¿Cuál es la estructura y la función de la proteína hipotética p284 de *Trypanosoma cruzi?*

6. OBJETIVOS

a. OBJETIVO GENERAL

- Predecir *In-silico* la estructura y función de la proteína hipotética >gi|74835222|sp|Q26940|Q26940_TRYCR de *Trypanosoma cruzi*

b. OBJETIVOS ESPECÍFICOS

- Encontrar similitudes significativas entre la secuencia de la proteína hipotética y las secuencias de las proteínas anotadas que se encuentran en las bases de datos
- Predecir *In-silico* las características Fisicoquímicas de la proteína hipotética
- Predecir *In-silico* las características de la estructura de la proteína hipotética (secundaria, terciaria y cuaternaria)
- Generar modelos en 3D de la proteína hipotética
- Predecir *In-silico* las características energéticas de la proteína hipotética

7. METODOLOGÍA

a. Detección de similitudes de la secuencia de la proteína p284 de Trypanosoma cruzi con las secuencias de proteínas anotadas:

- Se determinó la existencia de similitudes significativas entre la secuencia de la proteína hipotética y las proteínas con función ya determinada siguiendo los siguientes pasos:

1. Se buscó la secuencia en formato FASTA (que es una forma de representar las cadenas de aminoácidos en forma de texto) en la página de NCBI (National Center for Biotechnology Information)
2. La secuencia FASTA fue analizada por medio BLAST (Basic Local Alignement Search Tool) utilizando el algoritmo PSI-BLAST que busca similitudes en secuencias conservadas, y el algoritmo BLASTp que busca la similitud total. Los resultados de este análisis mostraron un *Score* o puntaje de similitud entre la proteína hipotética p284 y las proteínas con las que tuvo mayor coincidencia. También mostraron un *E-value* que es el valor de la significancia de la similitud, y por último un porcentaje de identidad. BLAST está conectado con CD search (Conserved Domain search) que mostró los dominios (que conforman la estructura terciaria) y homologías encontrados.

b. Predicción de las características fisicoquímicas de la Proteína Hipotética :

Para determinar las características Fisicoquímicas de la proteína hipotética se utilizó el programa *protparam*

c. Predicción de las características de la estructura secundaria y terciaria de la Proteína hipotética:

Para determinar la estructura secundaria de la proteína se utilizó el programa GOR IV por medio del cual se predijo el porcentaje de *láminas Beta y hélices alfa* que constituyen la proteína a partir de la secuencia de aminoácidos de la Proteína Hipotética en formato FASTA. Después por medio del programa SCOP (Structural Classification of proteins) se determinaron los motivos, o tipo de plegamientos de las secuencias conservadas de la proteína que definen la estructura terciaria por medio de *Motif Search*, y por último algunas características de la estructura cuaternaria fueron predichas por medio del análisis del programa Dockingserver.

d. Generación de modelos 3D de la proteína hipotética:

La generación de modelos 3D de la proteína se realizó por medio de la utilización de los programas SAS (Sequence Annotated by Structure), $(PS)^2$ y I-Tasser que permiten encontrar similitudes de la secuencia de la proteína hipotética en formato FASTA con las secuencias de estructuras 3D ya conocidas que se encuentran en *Protein Data Bank* (PDB). Por medio de estas comparaciones se creó un modelo de cómo se vería la proteína en tres dimensiones. Los retoques finales de la proteína se realizaron por medio de programas como *RasMol* y *Swiss PDB Viewer.*

e. Predicción de las características energéticas de la Proteína Hipotética:

Para la predicción de las características energéticas (energía de unión Enzima- sustrato y la superficie de interacción, etc) de la proteína se utilizó el programa Dockingserver.

8. RESULTADOS

8.1 Detección de similitudes de la secuencia de la proteína hipotética p284 de Trypanosoma cruzi con las secuencias de proteínas anotadas:

La búsqueda de similitudes para la secuencia de la proteína hipotética p284 se llevó a cabo por medio del programa BLAST. Los resultados indicaron que existen semejanzas significativas entre la secuencia de la proteína hipotética p284 y las proteínas de la *Superfamilia Smr (Small Mutations Related Domain)*, con una coincidencia de más de 200 residuos (figs 1 y 2 en rojo); también se encontraron similitudes de menor relevancia con las superfamilias KH-I y DUF1771 (figuras 1 y 2). La similitud entre las secuencias se mide por medio del *E-value*; entre menor sea este valor mayor es la significancia de la similitud entre las secuencias de las proteínas comparadas.

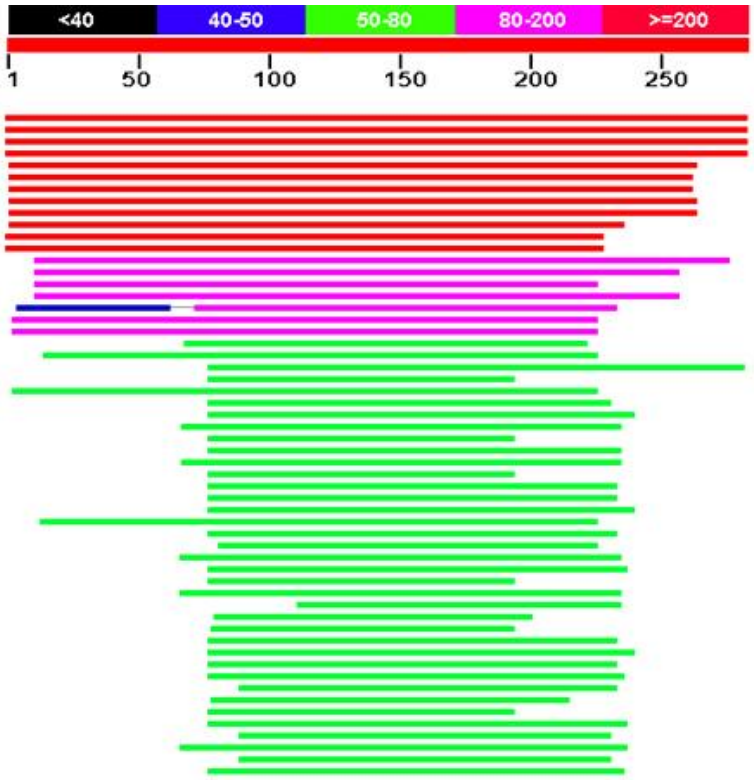


figura 1. Resultados de BLAST que muestran las coincidencias de la secuencia de la proteína hipotética con las superfamilias Smr (rojo). KH-I (verde) y DUF1771 (morado)

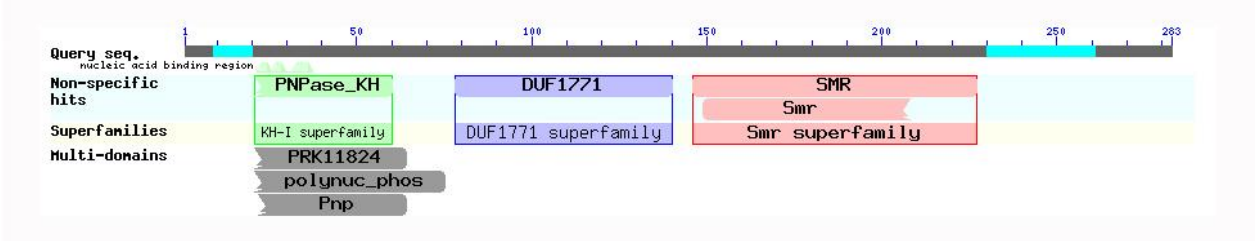


Figura 2. Resultados de BLAST que muestra las secuencias con las que la proteína hipotética p284tuvo más coincidencias

Tabla 1. Resultados de BLAST que muestran las coincidencias más significativas entre la proteína hipotética p284 de *T.cruzi* y otras proteínas

Descripción	Puntaje máximo	Puntae total	Cobertura	E-value	Id.max
hypothetical p284 protein [*Trypanosoma cruzi*]	572	572	100%	0.0	100%
hypothetical protein [*Trypanosoma cruzi* strain CL Brener]	521	521	100%	0.0	98%
hypothetical protein, conserved [*Trypanosoma cruzi*]	477	477	100%	2e-168	93%
hypothetical protein [*Trypanosoma cruzi* strain CL Brener]	471	471	100%	2e-167	93%
hypothetical protein, conserved [*Leishmania donovani*]	237	237	92%	4e-75	50%
hypothetical protein [*Leishmania braziliensis* MHOM/BR/75/M2904]	234	234	92%	2e-74	50%

8.2 Predicción de las características fisicoquímicas de la Proteína Hipotética:

La predicción se realizó por medio del programa Protparam.

8.2.1 Composición de aminoácidos:

Tabla 2. Resultados de Protparam que muestran la composición de aminoácidos de la proteína hipotética p284 de *T.cruzi*

Aminoácidos	Residuos	Porcentaje
Ala (A)	45	15.9%
Arg (R)	17	6.0%
Asn (N)	7	2.5%
Asp (D)	10	3.5%
Cys (C)	3	1.1%
Gln (Q)	12	4.2%
Glu (E)	27	9.5%
Gly (G)	29	10.2%
His (H)	7	2.5%
Ile (I)	11	3.9%
Leu (L)	18	6.4%
Lys (K)	22	7.8%
Met (M)	8	2.8%
Phe (F)	5	1.8%
Pro (P)	8	2.8%
Ser (S)	12	4.2%
Thr (T)	17	6.0%
Trp (W)	0	0.0%
Tyr (Y)	4	1.4%
Val (V)	21	7.4%
Pyl (O)	0	0.0%
Sec (U)	0	0.0%

Se encontró que el aminoácido más abundante de la proteína hipotética p284 es la alanina (Ala) con un total de 49 residuos, es decir un 15.9% del total de aminoácidos (tabla 2); la Alanina se clasifica como hidrofóbica y alifática. El segundo aminoácido más abundante es la Glicina (Gly) representando el 10.2% de los residuos; este aminoácido es neutral, alifático e hidrofóbico. En tercer lugar se encuentra el glutamato (Glu) representando el 9.5% de los aminoácidos, y es un aminoácido hidrofílico cargado negativamente.

8.2.2 Peso molecular:

El peso molecular calculado para la proteína Hipotética p284 fue de 299930.9 daltons lo que la clasifica como un polipéptido.

8.2.3 Carga:

El pI (pH al que la proteína tiene carga neta cero) calculado para la proteína hipotética fue de 8.33. También se determinó que la proteína tiene 39 residuos cargados positivamente (Arg + Lys) y 37 residuos cargados negativamente (Asp + Glu).

8.2.4 Composición Atómica:

Protparam calculó el número de átomos y la fórmula química de la proteína hipotética. Los valores se pueden observar en la tabla 3.

Tabla 3. Predicción del programa Protparam de la composición atómica y fórmula química de la proteína hipotética p284

Elemento		No. De átomos	Fórmula química
Carbono	**C**	1285	$C_{1285}H_{2119}N_{389}O_{410}S_{11}$
Hidrógeno	**H**	2119	
Nitrógeno	**N**	389	
Oxígeno	**O**	410	
Azufre	**S**	11	
Total átomos		4214	

8.2.5 Coeficiente de extinción:

Los resultados de los cálculos del coeficiente de extinción se pueden ver en la *tabla 4*.

Tabla 4. Coeficiente de extinción predichos por Protparam para la proteína hipotética p284 de *T.cruzi*

Coeficiente de extinción y absorbancia asumiendo todos los pares de residuos de cistínas	Coeficiente: 6085 Abs. 0.1% (=h/l): 0.203
Coeficiente de extinción y absorbancia asumiendo que todos los residuos de cisteína están reducidos	Coeficiente: 5960 Abs. 0.1% (=h/l): 0.199

8.2.6 Vida Media Estimada:

Tabla 5. Vida media estimada predicha por Protparam para la proteína hipotética p284 de *T.cruzi*

PARÁMETRO	VALOR
Vida media estimada	30 horas (reticulocitos de mamíferos, in vitro).
	>20 horas (levadura, in vivo).
	>10 horas (*Escherichia coli*, in vivo).

8.2.7 Índice de inestabilidad, Índice alifático y GRAVY:

Tabla 6. Valores del Índice de inestabilidad, GRAVY e Índice alifático predichos por Protparam para la proteína hipotética p284 de *T.cruzi*

Indice	Valor
Índice de inestabilidad	39.27, la proteína es estable
GRAVY (hidropaticidad)	-0.382 (hidrofílica)
Índice alifático	77.39

8.3 Predicción de las características de la estructura secundaria y terciaria de la Proteína hipotética:

Las características de la estructura secundaria de la proteína hipotética fueron predichas por medio del programa GOR IV.

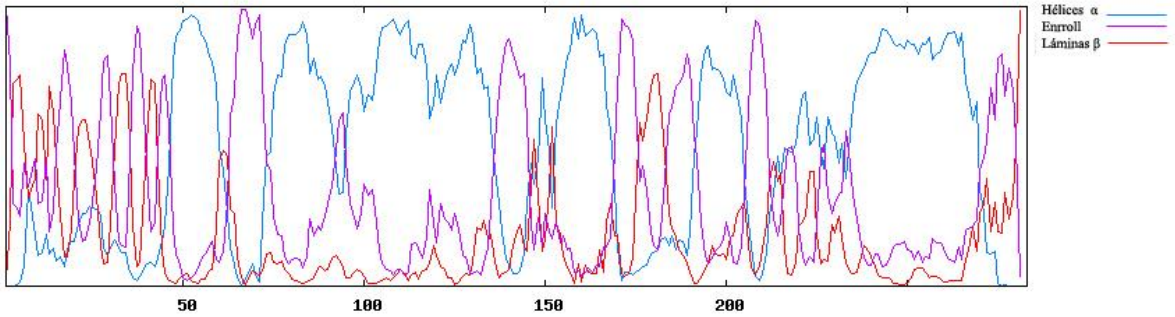


Figura 3. Representación gráfica de la estructura secundaria de la proteína hipotética p284 de *T.cruzi*

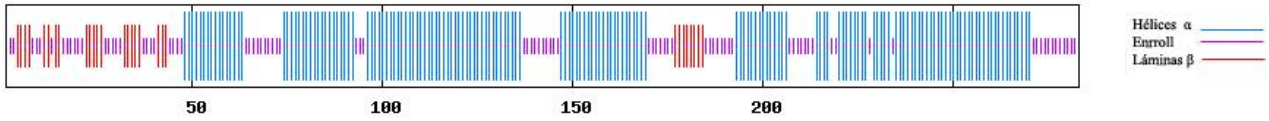


Figura 4. Representación gráfica de la estructura secundaria de la proteína hipotética p284 de *T.cruzi*

Por medio del programa *Motif Search* se determinaron los posibles motivos que tienen similitudes con la *proteína hipotética p284*.

Tabla 7. Motivos con mayor similitud en sus secuencias encontradas para la proteína hipotética p284 de *T.cruzi* predichos por el programa Motif Search

Motivo	Posición	PROSITE	Descripción
SMR	**149..230**	**PS50828**	**Perfil del dominio Smr**
KH_TYPE_1	1..59	PS50084	Type-1 KH domain profile.

Los resultados de SCOP (Structural Classification of Proteins) indican que el *dominio Smr* es de la clase α+β (ver tabla 8).

Tabla 8. Resultados de SCOP que muestran el tipo de plegamientos de la proteína hipotética p284 de *T.cruzi*

Plegamientos	Linaje	Superfamilias
IF3-like *beta-alfa-beta-alfa-beta(2); 2 capas; lámina mixta 1243, hebra 4 es antiparalela al resto*	Raíz: SCOP Proteínas α+β	**SMR domain-like**

8.4 *Generación de modelos 3D de la proteína hipotética:*

Para generar los modelos 3D de la proteína se utilizaron tres programas, I-TASSER, SAS y $(PS)^2$.

8.4.7 Resultados de I-TASSER:

Los resultados de I-TASSER muestran varios aspectos:

8.4.7.1 Modelo 3D generado por I-TASSER:

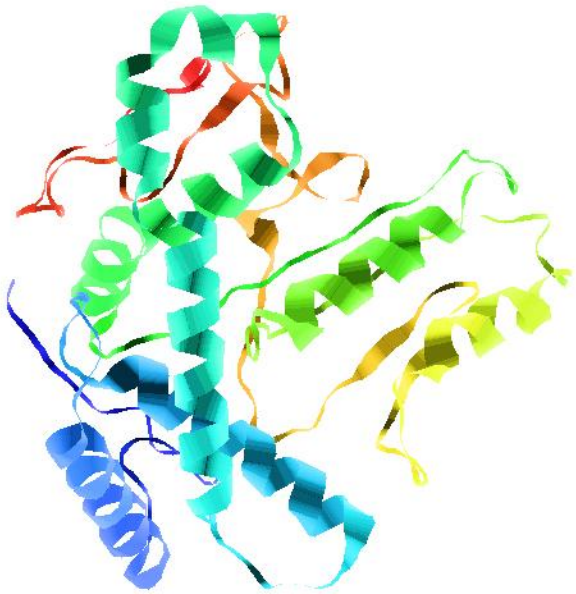


Figura 5. Modelo No 1. Generado por I-TASSER con el C-score más alto de -3.13

8.4.7.2 Top de plantillas relacionadas utilizadas:

Tabla 9. Top 10 de las plantillas utilizadas para la construcción de los modelos 3D de la proteína hipotética p284 de T.cruzi.

Rango*	PDB Hit	Iden 1	Iden 2	Cov.	Norm. Z-score	Descripción
1	2vkcA	0.30	0.12	0.36	4.20	**Dominio B3BP SMR, Hidrolasa, Mismatch Repair**
2	3b74A	0.08	0.18	0.89	1.27	Sec14 superfamily Proteína Señal, tráfico a través de la membrana
3	1uaaA	0.08	0.17	0.93	1.26	Helicasa, **hidrolasa** del DNA
4	2vkcA	0.29	0.12	0.36	1.47	**Dominio B3BP SMR, Hidrolasa, Mismatch Repair**
5	1lt7A	0.11	0.22	0.89	1.23	Transferasa, metil-transferasa
6	2vkcA	0.30	0.12	0.36	4.54	**Dominio B3BP SMR, Hidrolasa, Mismatch Repair**
8	2d9iA	0.27	0.10	0.33	0.79	**Dominio SMR, NEDD4-binding protein 2**
9	2x38A	0.08	0.17	0.96	0.40	Transferasa, phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit delta isoform
10	2vkcA	0.30	0.12	0.36	4.97	**Dominio SMR, Hidrolasa, Mismatch Repair**

***Iden 1:** porcentaje de identidad de la región alineada **Iden2:** porcentaje de identidad de toda la secuencia. Entre más alto el porcentaje mayor homología. Una identidad de secuencia que es alta en la región alineada y baja en el resto de la secuencia indica un motivo/dominio conservado presente tanto en la proteína hipotética como en la *proteína plantilla*. **Cov:** representa la cobertura del alineamiento y es igual al número de residuos alineados estructuralmente dividido por la longitud del modelo. **Norm.Z-score**: Un valor del Norm.Z-score refleja un alineamiento confiable. Si la mayoría de las plantillas tienen un Norm Z-score >1, la precisión del modelo final es usualmente alta.

8.4.7.3 Predicción de la función utilizando COFACTOR:

Tabla 10. Predicción de los números EC para la proteína hipotética p284 de *T.cruzi*

Rango	EC-score	PDB Hit	TM-score	RMSD[a]	IDEN[a].	COV.	Número EC
1	0.697	2gjxA **(Hidrolasa)**	0.330	5.96	0.067	0.647	**3.2.1.**52
2	0.690	1yq2A **(Hidrolasa)**	0.327	5.61	0.044	0.608	**3.2.1.**23
3	0.657	1o7aA **(Hidrolasa)**	0.331	5.71	0.039	0.636	**3.2.1.**52
4	0.623	2je8B **(Hidrolasa)**	0.250	6.52	0.043	0.530	**3.2.1.**25
5	0.604	1ea0A (Oxidorreductasa)	0.324	6.11	0.068	0.685	1.4.1.13

* **EC-score:** puntaje de confianza para la clasificación enzimática. **TM-score** es una medida de la similitud estructural global entre la secuencia query y la plantilla. **RMSD[a]** : es el RMSD entre residuos que están estructuralmente alineados por alineamiento TM .**IDEN[a]:** porcentaje de identidad de la secuencia en una región estructuralmente alineada. Entre más alto el porcentaje mayor homología **Cov:** representa la cobertura del alineamiento y es igual al número de residuos alineados estructuralmente dividido por la longitud del modelo

Tabla 11. Predicción de el sitio activo de la proteína hipotética p284 de *T.cruzi*

Rango	Cscore[LB]	PDB Hit	TM-score	RMSD[a]	IDEN[a]	Cov.	BS-score	Lig. Name
1	0.06	3b7zA (Prot.señal)	0.809	2.37	0.081	0.890	0.81	B7N
2	0.06	1oizB (Transporte)	0.496	5.12	0.058	0.739	0.66	TRT
3	0.06	1gjwA (transferasa)	0.422	5.54	0.074	0.657	0.64	PO4
4	0.06	1llwA (Oxidorreductasa)	0.420	6.62	0.032	0.749	0.66	F3S
5	0.06	3b74A (Prot.señal)	0.814	2.32	0.081	0.890	0.64	PEE
6	0.06	1jz6A (Hidrolasa)	0.402	5.52	0.061	0.629	0.68	DMS
7	0.06	1oipA (Transporte)	0.481	5.00	0.065	0.714	0.95	SO4
8	0.06	2wziA (Hidrolasa)	0.340	6.33	0.056	0.569	0.92	GOL

Cscore[LB]: valor de confianza de la predicción del sitio activo. De 0 a 1, entre más alto más confiable **TM-score** es una medida de la similitud estructural global entre la secuencia query y la plantilla. **RMSD[a]** : es el RMSD entre residuos que están estructuralmente alineados por alineamiento TM .**IDEN[a]:** porcentaje de identidad de la secuencia en una región estructuralmente alineada. Entre más alto el porcentaje mayor homología **Cov:** representa la cobertura del alineamiento y es igual al número de residuos alineados estructuralmente dividido por la longitud del modelo **BSscore:** medida de similitud local entre la sec. Query y la sec. plantilla.

Tabla 12. Ligandos predichos por I-TASSER y sus respectivas constantes de inhibición calculadas por Docking Server.

LIGANDO	Ki
B7N	No aparece
TRT	419.38 mM
PO4	14.86 mM
F3S	831.20 mM
PEE	No aparece
DMS	9.67 mM
SO4	4.28 mM
GOL	7.94 M

8.4.8 Resultados SAS:

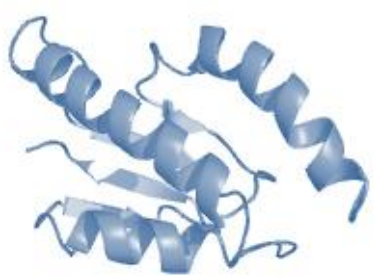


Figura 8. Aproximación de la estructura tridimensional de la proteína hipotética p284 obtenida de SAS basada en la estructura de la plantilla 2vkc que es una hidrolasa

La *tabla 13* muestra las plantillas utilizadas para la estructura propuesta. También se destaca el uso de proteínas con el dominio Smr o relacionadas.

Tabla 13. Resultados de los Hits estructurales detectados por SAS para la proteína hipotética p284 de *T.cruzi.*

Rango	Smith-Waterman score	% de Identidad	Superp a.a.	Long. sec.	*z*-score	*E*-value	PDB Hit	Nombre Proteína
1.	-	-	-	283	-	-	-	*Secuencia Problema*
2.	134	30.8%	91	96	169.3	0.014	2d9i:A	***Solución de la estructura del dominio smr de nedd4-binding protein 2***
3.	127	32.0%	103	105	160.4	0.045	2vkc:A	***Solución de la estructura del dominio b3bp smr***
4.	120	33.8%	77	81	153.7	0.1	3fau:B	***Estructura cristal del dominio relacionado con peq. Mutaciones en humanos (Smr)***
5.	115	37.1%	70	94	146.9	0.25	1x4m:A	*Solución de la estructura del dominio KH*

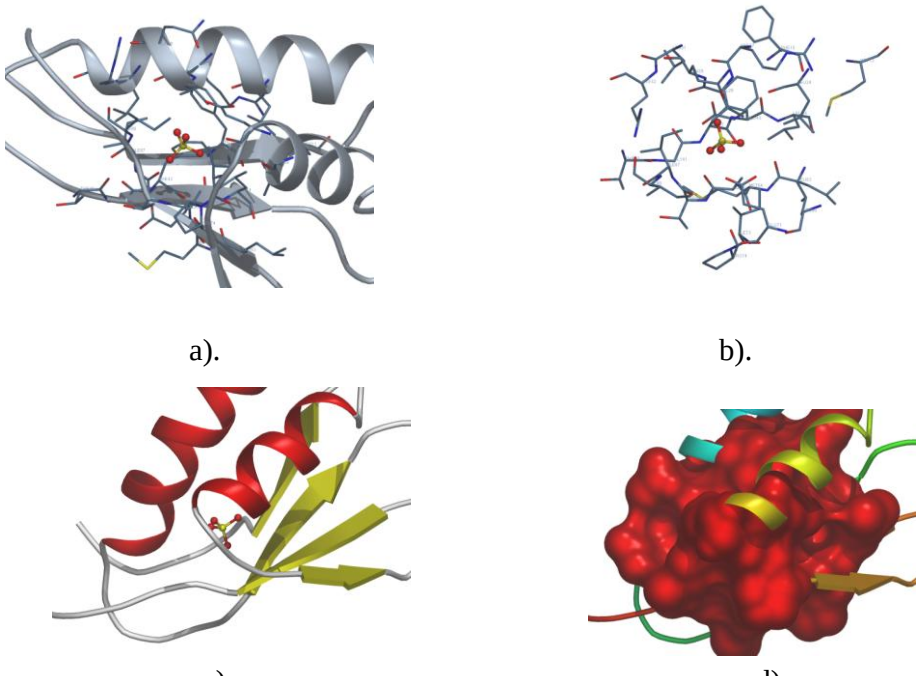
***Smith-Waterman score:** puntaje para alineamiento local. Detección de regiones similares entre dos secuencias **Z-score:** Valor estándar **E-value:** Expect Value

8.4.9 Resultados $(PS)^2$:

a). **b).**

Figura 9. a). Estructura predicha por $(PS)^2$ para la región 140 - 231 de la Proteína hipotética p284 de T.cruzi basada en la plantilla **b).** Factor B predicho por el programa $(PS)^2$

8.5 Predicción de las características energéticas de la Proteína Hipotética:



a). b).

c). d).

Figura 10. Geometría de la proteína hipotética p284 de *T.cruzi* generada por el programa Docking server **a).** Estructura general de la proteína hipotética p284 que **b).** Ligando SO_4 **c).** Estructura secundaria diferenciando hélices alfa (rojo) láminas beta (amarillo) y enrrollamientos aleatorios (blanco) **d).** superficie de la proteína

Tabla 14. Características energéticas de la geometría de la proteína hipotética p284 de *T.cruzi* con el ligando SO_4

Rango	Energía libre de unión estimada	Constante de Inhibición Ki	Energía vdW + Hbond + disuelta	Energía electrostática	Energía Intermolecular Total	Frecuencia	Superficie de Interacción
1.	-3.23 kcal/mol	4.28 mM	-3.22 kcal/mol	-0.01 kcal/mol	-3.23 kcal/mol	90%	195.913

Tabla 15. Resultados de Docking server en donde se muestran interacciones de la proteína con el ligando

Ligando	Proteína
O	VAL28
S	VAL28
S	PHE42
O	PHE42
S	ALA85
O	ALA85
S	ILE87
O	ILE87

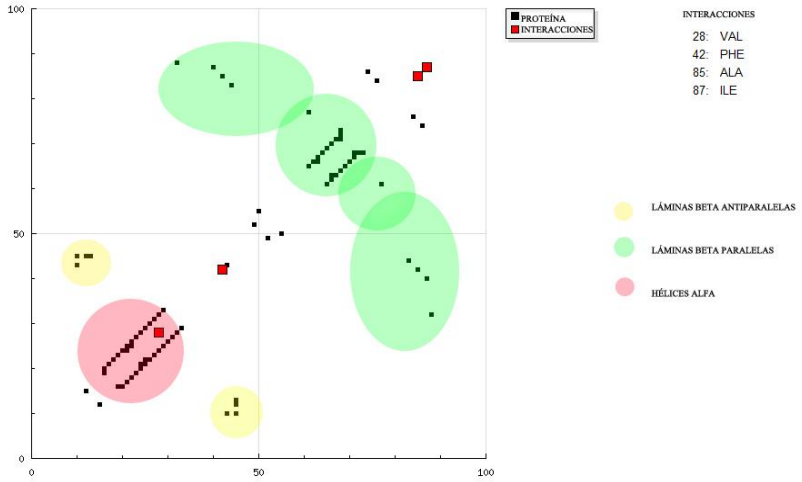


Figura 16. Gráfica HB plot que muestra las interacciones de la proteína hipotética p284 generada por Docking server.

9 DISCUSIÓN

El análisis de la secuencia de la proteína hipotética p284 de *T.cruzi* a través de los diferentes programas bioinformáticos permitió la obtención resultados concluyentes. Para empezar, los análisis realizados con BLAST mostraron que esta proteína posee el *dominio Smr* (Small mutations related domain), que es el dominio característico de un grupo de proteínas denominadas *MutS2*, homólogas de las proteínas *MutS1*, que son enzimas altamente conservadas comprometidas con las reparaciones de los desajustes del ADN, la recombinación meiótica y otras modificaciones genéticas (23, 24). Las proteínas homólogas de MutS han sido encontradas en muchas especies incluyendo eucariotas, Archaeas y bacterias, y han sido agrupadas dentro de la *familia MutS* (23,25). Las proteínas *MutS2* carecen del dominio de reconocimiento de *MutS1*, pero contienen una extensión *C-terminal* adicional: *Small MutS-related (Smr) domain* (23, 24, 25). Las proteínas MutS2 reconocen las transiciones entre el ADN de una sola hebra (ssDNA) y el ADN de doble hebra (dsDNA) que ocurre durante la recombinación (23, 25). Mientras que las proteínas MutS1 reconocen y remueven los errores de coincidencia heterólogos, las proteínas MutS2 parecen interferir en las recombinaciones tanto heterólogas como homólogas (23). De este modo las proteínas MutS2 pueden regular la reorganización del ADN endógeno en el entrecruzamiento meiótico y en la segregación cromosómica, así como en la incorporación de ADN exógeno (23, 24, 25). Los resultados de BLAST también mostraron que la secuencia analizada está altamente conservada entre diferentes cepas de *T.cruzi*, y también en otros kinetoplástidos como *Leishmania* (ver tabla 1).

Por otro lado, las características fisicoquímicas predichas por el programa Protparam revelaron varios aspectos muy importantes acerca de la proteína hipotética. Su secuencia tiene 283 residuos, 39 de ellos están cargados positivamente y 37 están cargados negativamente. Esta información es importante porque se ha detectado que la composición de aminoácidos varía sutilmente entre proteínas con diferentes localizaciones subcelulares (26). En un pH neutro las proteínas intracelulares tienden a tener una mayor fracción de residuos cargados negativamente mientras que las proteínas nucleares se caracterizan por tener un porcentaje relativamente elevado de aminoácidos cargados positivamente (26, 27). La carga positiva de estas proteínas les sirve para unirse al ADN, que está cargado negativamente (26, 27) Según lo anterior es probable que la proteína hipotética p284 sea una proteína nuclear pues tiene mayor cantidad de residuos cargados positivamente que de residuos cargados negativamente.

Otro valor calculado por este programa fue el del peso molecular de la proteína. Este valor es importante porque los polímeros de aminoácidos se pueden diferenciar de acuerdo a su peso; las moléculas con pesos moleculares que van de varios miles a millones de daltons se llaman *polipéptidos* y aquellas con pesos moleculares bajos, típicamente consistiendo de menos de 50 aminoácidos se llaman *péptidos* (27, 28). Debido a que el peso calculado para la proteína hipotética sobrepasa los cientos de miles de daltons (299930.9) esta se puede clasificar como un polipéptido. Se debe tener en cuenta que debido a que en el cálculo que realiza el programa para predecir el peso molecular de la proteína no se incluyen las modificaciones post-translacionales (glicosilaciones) el valor del peso molecular no es preciso, sin embargo es una buena aproximación al peso real, que solo puede ser determinado por medio de una espectroscopia de masas realizada en un laboratorio (16). Protparam también determinó el carácter básico o ácido de la proteína por medio del cálculo del *punto isoeléctrico teórico (pI)* que es el pH en el que la proteína tiene igual número de cargas positivas y negativas y por lo tanto la carga neta es cero (27, 28). Todas las proteínas tienen un grupo amino en un extremo y un grupo carboxilo al otro al igual que numerosas cadenas de aminoácidos, algunas del las cuales están cargadas; por esta razón, cada proteína tiene una carga neta (11, 27, 28). La carga neta de una proteína está fuertemente influenciada por el pH de la solución (11, 28). Cuando el valor del pH baja hay más iones H+ en la solución, estos iones de H+ se unen a los sitios negativos en los aminoácidos volviendo positiva la carga de la proteína; de manera opuesta, cuando el valor del pH se hace más básico la proteína se hace más negativa (27, 28, 29). En el caso de la proteína hipotética el pI teórico fue de 8.33, es decir que la proteína es de carácter básico. Las proteínas básicas se asocian con funciones que tienen que ver con el ADN como en el caso de las histonas (26, 29).

Otro aspecto fisicoquímico determinado por Protparam fue el *coeficiente de extinción* que indica cuanta luz es absorbida por una proteína en una longitud de onda determinada (11, 30). Generalmente el coeficiente es calculado por medio de una espectrofotometría de rayos UV pero también puede ser calculada utilizando la composición de aminoácidos y la absorbancia de la cisteína, la tirosina y el triptófano por medio de una ecuación (11, 30). En los resultados del coeficiente de extinción siempre se obtienen dos valores, uno teniendo en cuenta todos los pares de residuos de cisteínas y otro asumiendo que todos los residuos de cisteína están reducidos. Esto se debe a que el grupo *sulfhidrilo* de la cisteína es altamente reactivo siendo su reacción más común una oxidación reversible que forma un *disulfido* (30). La oxidación de dos moléculas de cisteína forma *cistina*, una molécula cuyo enlace produce la fomación de *puentes disulfuro* que ayudan a estabilizar muchos polipéptidos y proteínas

(27, 28, 30). El coeficiente de extinción para la proteína hipotética p284 teniendo en cuenta todos los pares de residuos de cisteínas fue de 6085 y asumiendo que todos los residuos de cisteína están reducidos es de 5960 (ver tabla 4). Estos resultados pueden tener hasta un 10% de error debido a que la proteína no tiene triptófano (16). El hecho de que la proteína no tenga triptófano corrobora la idea de que la proteína hipotética p284 es una proteína nuclear pues estas proteínas tienen generalmente un bajo porcentaje de residuos aromáticos (30). Un compuesto con un alto valor de coeficiente de extinción molar es muy eficiente en la absorción de luz de la longitud de onda adecuada y, por lo tanto, puede detectarse por medidas de absorción cuando se encuentra en disolución a concentraciones muy bajas (30).

Protparam también calculó la *vida media estimada* de la proteína que es una predicción del tiempo que le toma a la mitad de la cantidad de una proteína para desaparecer después de ser sintetizada en la célula. Los valores del tiempo estimado que duraría la proteína hipotética p284 en diferentes medios biológicos (in vivo e in vitro) obtenidos por este programa indican que es estable (ver tabla 5).; la vida media de las proteínas es altamente dependiente de la presencia del amino ácido N-terminal y por lo tanto de la estabilidad total de la proteína (16). La importancia de los residuos del N-terminal se conoce generalmente como la *regla de la N-terminal* que en resumen dice que el aminoácido que se encuentre en el N-terminal determina la vida media de las proteínas (16, 28). Los valores de la vida media estimada para los diferentes aminoácidos han sido calculados en mamíferos, levadura y *E.coli*, a partir de estos valores que se estimó cual sería la vida media para la proteína hipotética p284 de acuerdo del aminoácido que posee en el extremo N-terminal. Existen otros dos valores calculados por este programa que indican que la proteína es estable, estos son el *índice de inestabilidad* y *el índice alifático.* El *índice de inestabilidad* provee un estimado de la estabilidad de la proteína en un tubo de ensayo; una proteína que obtenga un valor de inestabilidad menor a 40 se considera como estable, y mayor a 40 como inestable (16). La proteína hipotética p284 de *T.cruzi* obtuvo un valor de 39.27 y es considerada como estable (ver tabla 6). El *índice alifático,* que es el volumen relativo ocupado por cadenas alifáticas (Alanina, Valina, Isoleucina y Leucina), obtuvo un valor alto (ver tabla 6) indicando que la proteína tiene un alto contenido de cadenas alifáticas que le proveen termorresistencia y por lo tanto estabilidad en diferentes condiciones (11, 16, 28). Por último este programa calculó el valor GRAVY (Grand Average of Hidrophaty) que permitió determinar que la proteína es hidrofílica pues el valor resultante fue negativo (ver tabla 6). La presencia de grupos polares en las cadenas laterales localizadas en la superficie de las proteínas, a diferencia de los grupos

hidrofóbicos que se encuentran al interior, es la que favorece la solubilidad de las proteínas en el agua, y por lo tanto la que determina el valor GRAVY (16).

Por medio del programa GOR IV se predijeron los patrones de plegamiento de la estructura secundaria de la proteína hipotética. Los arreglos regulares de la estructura secundaria son de dos tipos principales: hélices alfa, con patrones de repetición de puentes de hidrógeno locales y láminas beta con patrones repetitivos entre partes distantes de la cadena polipeptídica (11, 27, 28)Los resultados de la predicción mostraron que 116 de los residuos de la proteína, es decir el 58.66%, forman hélices α, 88 de los residuos, es decir el 31.10%, forman enrrollamientos aleatorios y 29 de los residuos, es decir el 10.25%, forman láminas β (figs 3 y 4). ,

En cuanto al análisis de los motivos de la proteína y dominios de la proteína que constituyen la estructura terciaria se encontraron los siguientes resultados. Por medio del software Motif.search se determinó que la secuencia de la proteína hipotética tiene una gran similitud con el motivo SMR *de la superfamilia Smr* coincidiendo con los resultados de BLAST (ver tabla 7). Por medio del programa SCOP (Structural classification of proteins) la proteína hipotética fue clasificada dentro de la clase de las proteínas α+β pues contiene tanto hélices α como láminas β, principalmente láminas beta antiparalelas y regiones segregadas de alfa y beta (ver tabla 8). SCOP también clasificó la proteína hipotética dentro del grupo de las proteínas que se pliegan de la forma *IF3-like* que se describe en detalle en la tabla 8.

La estructura tridimensional de la proteína fue determinada en primer lugar por I-TASSER que generó cinco modelos a partir de partida estructuras "plantilla" (con secuencias similares) obtenidas del *Protein Data Bank*. Cada uno de los modelos generados estaba acompañando por un *C-score* que es el valor de confianza de la predicción de la estructura tridimensional. El *C-score* es un estimado de la calidad de los modelos predichos y es calculado basándose en la significancia (*Z-score*) de los alineamintos hechos automáticamente en el programa LOMETS, y la convergencia (densidad de clusters) de las simulaciones de I-TASSER (31). Los puntajes del *C-score* se encuentran típicamente entre -5 y 2, en donde un puntaje más alto refleja un modelo de mejor calidad. En general, los modelos con un *C-score* > -1.5 tienen plegamientos (31). En el caso de la proteína hipotética p284 de *Trypanosoma cruzi* el modelo 1 (figura 5) con el mejor C- score (-3.13) se considera dentro del promedio pero no alcanza valores mayores a -1.5, indicando que puede que la predicción tenga algunos plegamientos que no están del todo correctos, sin embargo es un buen C-*score* que indica que hay buena aproximación a la estructura tridimensional de la proteína. En la tabla 9 se puede ver el top

de las plantillas utilizadas por I-TASSER para la construcción del modelo tridimensional; estos resultados hacen evidente que la construcción de los modelos tridimensionales realizada por I-TASSER se basó en proteínas con el *dominio Smr*, pues 5 de las 10 plantillas utilizadas tienen este dominio (ver tabla 9 en rojo).

I-TASSER también predijo la posible función de la proteína utilizando el valor de el EC- score que es un puntaje de confianza que puede ser usado para la anotación funcional de la proteína (31). Un EC-score > 1.1 es un buen indicador de similitud en la función. A pesar de que los valores del EC-Score no llegan a 1.1 existe una coincidencia en los 3 primeros tres números de los valores EC (ver tabla 10 resaltado en rojo) que puede indicar un grado de alto de confianza en la predicción (31). La mayoría de los hits de PBD tienen función de hidorlasas lo que indica que es probable que la protreína hipotética p284 tiene esta función. Este programa también predijo los posibles ligandos de la proteína hipotética que se pueden ver en la tabla 11.

Otro de los programas que se utilizó para la predicción de la estructura 3D de la proteína hipotética fue SAS (Sequence Annotated by Structure) que es una herramiena para encontrar información estructural de una secuencia dada utilizando la secuencia FASTA de una proteína para escanearla contra todas las proteínas de estructura 3D conocidas en el Protein Data Bank (PDB). Este programa es un servidor automático de modelación de homologías. Utiliza un consenso efectivo combinando el uso de los programas PSI-BLAST, IMPALA, y T-Coffee. La estructura tridimensional es generada utilizando el programa MODELLER (Figura 9); después de superponerla con la estructura generada por I-TASSER en el programa RasMol se evidenció que son muy similares. Por último, el programa $(PS)^2$ (figura 9) realizó otra predicción de la estructura tridimensional muy similar a la de I-TASSER y SAS. Las comparaciones entre las estructuras generadas por los tres programas fueron realizadas utilizando RasMol y Swiss pdb viewer. Adicionalmente $(PS)^2$ generó una gráfica que muestra el factor B de la secuencia de la proteína (figura 9). El factor B describe la reducción de la intensidad de dispersión coherente de rayos X, neutrones o electrones por un cristal debido al movimiento térmico de los átomos en la red cristalina (32). La teoría dice que este factor puede indicar donde hay errores en la posición de los átomos en la construcción de los modelos tridimensionales (32).

Por último los resultados de características energéticas y de unión al ligando de la proteína hipotética p284 generados por medio del programa Docking server se realizaron utilizando como ligando al SO_4. que fue el ligando predicho con I- TASSER con la constante de inhibición *Ki* más baja, es decir con mayor afinidad con la proteína hipotética (ver tabla 12). La estructura tridimensional utilizada para el

Docking fue obtenida a partir del modelo generado por I-TASSER con el mayor C-score con algunas modificaciones menores realizadas en RasMol teniendo en cuenta los modelos de (PS^2) y SAS. Con el archivo .pdb de la estructura y la conformación del ión SO_4 obtenida de *Pubchem* Docking server generó imágenes de la geometría de la proteína hipotética p284 y de la posible conformación del sitio activo (ver figura 10) (33). En la tabla 14 se pueden observar los valores de la energía de unión entre la proteína y el ligando calculados por este programa, en primer lugar se presenta el valor de la *Energía libre de unión estimada* que indica si la unión de la proteína con el ligando ocurre de manera espontánea o no. Si este valor, también conocido como ΔG, es negativo indica que la reacción ocurre de manera espontánea (exergónica), si es positivo indica que la reacción no es espontánea (endergónica), y si es igual a cero indica que está en equilibrio (11). En este caso el valor de la energía libre de unión es negativa (-3.23 kcal/mol) lo que indica que la reacción ocurre de manera espontánea. El segundo valor en la tabla es la *Constante de inhibición Ki* que es la concentración de ligando necesaria para ocupar la mitad de los sitios de enlace al ligando disponibles; mientras más afín es una proteína a un ligando, menor cantidad de ligando requiere para ocupar el 50% de sitios de unión a ligando (11, 33). El valor de la *Ki* para la interacción de la proteína hipotética p284 con el ligando SO_4 fue de 4.28 mM, el valor más bajo de todas las constantes de inhibición calculadas por Docking Server para los 8 ligandos predichos por I-TASSER. Este valor bajo indica que hay una muy buena interacción proteína-ligando (33). El tercer valor en la tabla es *Energía vdW + Hbond + disuelta* y nos muestra la sumatoria de 3 fuerzas energéticas, las *interacciones de Van der Waals*, los *puentes de hidrógeno* y la *energía de desolvación*. Las *interacciones de Van der Waals* se refieren a las fuerzas de repulsión o atracción entre moléculas y definen el carácter químico de muchos compuestos orgánicos. Los *puentes de hidrógeno* son enlaces formados por la atracción de átomos electronegativos y átomos de hidrógeno y La *energía de desolvatación* es el cambio en la energía libre cuando se remueve el solvente del medio. La suma de estas tres fuerzas fue de -3.22 kcal/mol. El cuarto valor de la tabla es la *Energía electrostática* que muestra el grado de atracción o repulsión electrostática (cargas eléctricas en reposo) de la unión de la proteína y el ligando. La unión de la proteína hipotética p284 con el ligando SO_4 hace que esta adquiera una carga electrostática negativa de -0.01 kcal/mol. El quinto valor de la tabla es la *Energía intermolecular total* que coincide casi exactamente con el valor de la energía libre de unión (-3.23) y es la energía de la sumatoria de las fuerzas de Van der Waals y los puentes de hidrógeno de las moléculas. El sexto valor de la tabla es la Frecuencia de la ocurrencia de la interacción proteína-ligando. En el caso de la interacción de la proteína hipotética p284 y el ligando SO_4 es del 90%, un porcentaje bastante alto (33). Por último la

tabla 14 muestra la *Superficie de interacción* entre el ligando y la proteína (195.913Å). Entre mayor sea la superficie de interacción mejor será la interacción entre la proteína y el ligando (35,30).

Las interacciones específicas entre los residuos del sitio activo y el ión SO_4 se muestran en la tabla 15. Los aminoácidos del sitio activo son Valina, Fenilalanina, Alanina e Isoleucina y los tres interactúan con el SO_4.

Por último este programa generó un gráfico HB plot (Figura 11) que muestra la relación entre la estructura de la proteína y su comportamiento dinámico. El *HB plot* ofrece un método sencillo para analizar la estructura secundaria y terciaria de una proteína (32, 33). Los puentes de hidrógeno que estabilizan los elementos de la estructura secundaria y los que se forman entre residuos de aminoácidos distantes pueden ser distinguidos en el HB plot, y por lo tanto los aminoácidos involucrados en la estabilización de la proteína y su función pueden ser identificados (33). Esta es una forma gráfica de representar las interacciones que ocurren en la proteína (32, 33).

Conclusiones

Los resultados de esta investigación permitieron concluir que la proteína hipotética p284 es una proteína nuclear de la familia Muts2 que probablemente actúa como una hidrolasa en tareas que tienen que ver con la recombinación, y reparación del ADN de *Trypanosoma cruzi*. Su estructura tridimensional es muy similar a la del dominio SMR que es el dominio característico de esta familia. En cuanto a sus características energéticas se encontró que la proteína tiene una afinidad importante con el ión SO_4, que fue el ligando predicho por el programa I-TASSER con el menor valor de la constante de Inhibición Ki. También se encontró que la secuencia de la proteína está altamente conservada entre los diferentes tipos de *Trypanosoma cruzi* y otros kinetoplástidos como Leishmania.

BIBLIOGRAFÍA CITADA

1. Levin M. Contribución del proyecto genoma de *Trypanosoma cruzi* a la comprensión de la patogénesis de la cardiomiopatía chagásica crónica. Medicina (Buenos Aires) 1999; **59**: 18-24
2. De Pablos LM. Análisis global de la familia multigénica MASP (Mucin Associated Surface Proteins) de *Trypanosoma Cruzi*. Primera Edición. Editorial de la Universidad de Granada. Granada., España. 2010, 230 p.
3. Imbert J, Figueroa A, Gómez J. Tripanosomiasis americana o "mal de Chagas" otra enfermedad de la pobreza. Elementos 2003; **49**: 13-21.
4. Lubec G, Afjehi-Sadat L, Yang J, Pradeep J. Hypothetical proteins: theory and practice based upon original data and literature. Progress in Neurobiology 2005; **77**: 90-127

5. Zarembisnki TI, Hung LW, Mueller-Dieckmann HJ, Kim KK., Yokota H, Kim R, Kim SH. Structurebased assignment of the biochemical function of an hypothetical protein: a test case of structural genomics. Proc. Natl. Acad. Sci. USA, 1998; **95**: 15189-15193

6. Dias JC, Silveira AC, Schofield CJ. The Impact of Chagas Disease Control in Latin America - A Review Mem Inst Oswaldo Cruz. . Rio de Janeiro., Brazil. 2002; **97** (5): 603-612

7. Moncayo Á, Silveira, A. Current epidemiological trends for Chagas disease in Latin America and future challenges in epidemiology, surveillance and health policy Mem Inst Oswaldo Cruz, Rio de Janeiro., 2009 .**104** (I): 17-30

8. Ellis T, Morrison A. Application of bioinformatics on parasitology. International Journal for parasitology, 2005. **35** (5): 463-464.

9. Barreto E. Bioinformática: Una oportunidad y un desafío. Revista Colombiana de Biotecnología, 2008. **10** (1): 132-138.

10. Brock O, Bunette TJ. Predicting Protein Structure with Guided Conformation Space Search. Technical report. University of Massachusetts, http://www.redaktion.tu-berlin.de/fileadmin/fg170/Publikationen_pdf/tr05-63_01.pdf consultado el 25 de febrero de 2012

11. Cozzone, AJ. Fundamental Chemical Properties Institute of Biology and Chemistry of Proteins Encyclopedia of life sciences, Macmillan Publishers Ltd, Nature Publishing Group. Lyon., Francia. 2002, 20 p.

12. Flórez L. BLAST en detalle. Bioinfórmate. http://bioinformate.uniandes.edu.co/cap6.html Universidad de los Andes, Bogotá Colombia. Capítulo 6. consultado el 20 de octubre de 2011.

13. Marchler-Bauer A et al. CDD: a Conserved Domain Database for the functional annotation of proteins, Nucleic Acids Res. 2011; **39** (D): 225-9.

14. Marchler-Bauer A et al. CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res 2009; **37** (D): 205-10.

15. Marchler-Bauer A, Bryant SH (2004), CD-Search: protein domain annotations on the fly.", Nucleic Acids Res 2009; **32** (D): 327-331.

16. Walker JM. Protein Identification and Analysis Tools on the ExPASy Server. The Proteomics Protocols Handbook. Humana Press .2005, 607p.

17. Garnier J, Gibrat JF, Robson B. GOR method for predicting protein secondary structure from amino acid sequence. Methods Enzymol 1996; **266**:540-562

18. Kyoto University Bioinformatics Center. GenomeNet. http://www.genome.jp/en/. Consultado el 25 en enero de 2012.

19. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol 1995. **247**: 536-540

20. Yang Zhang. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics, 2008. **9** (40): 1-20

21. Chen CC, Hwang JK, Yang JM. $(PS)^2$: protein structure prediction server. NAR, 2006, **34**: W152-W157

22. Milburn D, Laskowski RA, Thornton JM. Sequences annotated by structure: a tool to facilitate the use of structural information in sequence analysis. Prot. Eng. 1998; 11, 855-859

23. Diercks T, Eiso AB, Daniels MA, de Jong RN, Besseling R, Kaptein R, Folkers GE. Solution Structure and Characterization of the DNA-Binding Activity of the B3BP–Smr Domain. Journal of Molecular Biology 2008. **383** (5): 1156–1170

24. Eisen, JA. A phylogenomic study of the MutS family of proteins. Nucleic Acids Research 1998; **26** (18): 4291–4300

25. Fukui1 K., Kosaka1 H, Kuramitsu S. Masui1 R. *Nuclease* activity of the MutS homologue MutS2 from *Thermus thermophilus* is confined to the Smr domain. Nucleic Acids Research 2007. **35**, (3): 1-20

26. Andrade, MA. Adaptation of protein surfaces to subcellular location, J Mol Biol. 1998. **276**: 517-525

27. Oxford University press. Biochemistry in perspective. Amino acids, peptides and proteins, chapter 5 http://oup.com/us/pdf/mckee/ch5p123-143. consultado el 5 de marzo de 2012.

28. Franco L. Enzimas: qué son y para qué sirven. Rev.R.Acad.Cienc.Exact.Fís.Nat 2007; **101** (2): 399-417

29. Acosta KY, Zavala JE. Proteínas de unión a DNA. Rev Biomed 1996; **7**:163-172

30. Thermo Scientific. Extinction Coefficients: A guide to understanding extinction coefficients, with emphasis on spectrophotometric determination of protein concentration. http://www.piercenet.com/files/TR0006-Extinction-coefficients.pdf citado el 3 de marzo de 2012

31. Ambrish R, Alper K, Yang Z. I-TASSER: a unified platform for automated protein structure and function prediction. Nature Protocols. 2010; **5**: 725-738

32. Karadaghi SA. Introduction to protein structure and structural bioinformatics; http://www.proteinstructures.com/Structure/Structure/proteinstructure-databases2.html citado el 20 de enero de 2012.

33. Leis S, Schneider S, Zacharias M. *In Silico* Prediction of Binding Sites on Proteins. Current Medicinal Chemistry. 2010; **17***:* 1550-1562