# Universidad Nacional Autónoma de México

## Facultad de Estudios Superiores Acatlán

Identificación asistida por computadora de dominios funcionales y su evolución en proteínas

# T E S I S

QUE PARA OBTENER EL TÍTULO DE:

Licenciado en Matemáticas Aplicadas y Computación

PRESENTA:

Ramón Arnulfo Flores Rodríguez

TUTOR

Dr. Fidel Alejandro Sánchez Flores


Junio, 2017

# Agradecimientos

A Dios, pues estoy seguro que sin su gracia y bendiciones este trabajo no habría sido posible.

A mis padres Mauricio y Honorina, que creyeron en mí desde antes de comenzar este viaje escolar y que me apoyaron en todo lo que pudieron. Que este trabajo los haga sentirse orgullosos del resultado de todos sus esfuerzos.

Al Dr. Fidel Alejandro Sánchez Flores por la oportunidad de participar en un proyecto de investigación en uno de los mejores institutos de la UNAM. Por el apoyo brindado durante mi estancia en la ciudad de Cuernavaca y el conocimiento que compartió conmigo para entender el alcance del proyecto.

A la Dr. Alejandra Escobar, Dr. Ernestina Godoy, Mtra. Verónica Jiménez, Dr. Luciana Raggi, Dr. Leticia Vega, Mtro. Jerome Verleyen y Lic. Karel Estrada, miembros de la Unidad de Secuenciación Masiva y Bioinformática, quienes colaboraron con sugerencias y correcciones al trabajo.

A los profesores de la FES Acatlán: Beatriz Herminia de Guadalupe Arreola Ramírez, Ricardo Domínguez Esquerro y Adriana Dávila Santos.

A los excelentes profesores y revisores de este trabajo, Eduardo Loza Pacheco, Liliana Gutiérrez Flores, Andrés Hernández Balderas y Mahil Herrera, que con sus sugerencias hicieron de este un mejor trabajo.

*Y todo lo que hagan, háganlo de corazón, como para el Señor y no como para la gente. Colosenses 3:23*

# Glosario de términos

**Anotación** Es una descripción de una proteína o un fragmento de proteína en términos de un tema tal como una función, enfermedad, estructura, etc.

**Base nitrogenada** Compuesto orgánico cíclico con uno o más átomos de nitrógeno. Son los componentes del ADN y ARN (adenina, guanina, citosina, timina/uracilo).

**Dominio funcional** Es un fragmento de proteína que se caracteriza por brindarle una función a la proteína en que se encuentra.

**Filogenia** Estudio de la relación o parentesco entre entes biológicos a partir de un ancestro común.

**Genoma** Conjunto de bases nucleótidas que representa todos los genes de un organismo.

**Secuenciación** Hace referencia a la identificación del orden de nucleótidos o aminoácidos en una cadena de ADN o en una proteína, respectivamente, a través de una serie de métodos y técnicas bioquímicas.

**Transcriptoma** Colección de todas las "lecturas" de genes presentes en una célula.

# Resumen

La identificación de dominios funcionales en proteínas es una técnica que permite identificar posibles relaciones evolutivas entre su estructura y función. Algunos de los métodos actuales para la identificación de dominios en proteínas aunque son eficientes están implementados en herramientas web que dependen de factores como concurrencia, disponibilidad y velocidad de red, aspectos en su mayoría no controlados por los usuarios.

En el presente trabajo se describe la creación de una herramienta de análisis bioinformático llamada SDA (*Scan Domain Architecture*) que en contraste a las actuales, solo requiere conexión a Internet para su instalación, y que además de identificar los dominios funcionales muestra su arquitectura para una o más secuencias de proteínas. Como una ventaja, esta herramienta permite la búsqueda de arquitecturas de dominios en archivos tabulares a manera de datos "locales". El desarrollo del *pipeline* está hecho en Perl debido a su facilidad de uso, además de ser uno de los tres lenguajes preferidos por los bioinformáticos[1].

En el capítulo primero se abordan los fundamentos bioquímicos y bioinformáticos para comprender de mejor manera el alcance y logros del desarrollo. El segundo capítulo describe algunas de las herramientas actuales que se usan para la identificación de dominios funcionales, se plantea la hipótesis del trabajo y se detalla el objetivo general así como los objetivos particulares para las pruebas de SDA. El capítulo tercero plantea los métodos del trabajo, se mencionan las herramientas usadas, la definición de la métrica de similitud para arquitecturas y el desarrollo de SDA como herramienta alternativa a las actuales. En el cuarto capítulo se muestran los resultados obtenidos con SDA para los objetivos particulares planteados anteriormente.

En las pruebas, SDA mostró obtener mejores resultados en un menor tiempo posible que con las demás herramientas utilizadas. Además de contar con la ventaja de permitir al usuario buscar arquitecturas de dominios en archivos locales, lo que implica un ahorro considerable de tiempo.

En las conclusiones se hace notar la efectividad de SDA, así como la respuesta a la hipótesis planteada y al objetivo tanto general como particulares. Por último, se planea expandir las capacidades de SDA para mejorar así la anotación y curación manual de proteínas.

# Índice general

# Capítulo 1

# Marco teórico

La bioinformática es un campo de investigación cuya existencia se remonta a principios de los 60s, en la que los biólogos comenzaron a crear modelos para tratar de explicar las interacciones entre distintas moléculas[2][3] y comprender cómo es que las proteínas se conforman para generar una función[4]. Uno de los resultados más importantes de este campo es la secuenciación del genoma humano, que ayudó a entender mejor distintos aspectos del cuerpo humano[5][6][7].

La primer secuenciación de una proteína completa fue realizada en la década de 1945-1955 por Frederick Sanger y sus colaboradores[8]. Su trabajo fue tomado como referencia para desarrollar métodos más efectivos para identificar los componentes de las proteínas, lo que propició un crecimiento en la cantidad de secuencias conocidas, ilustrado en la Figura 1.1.

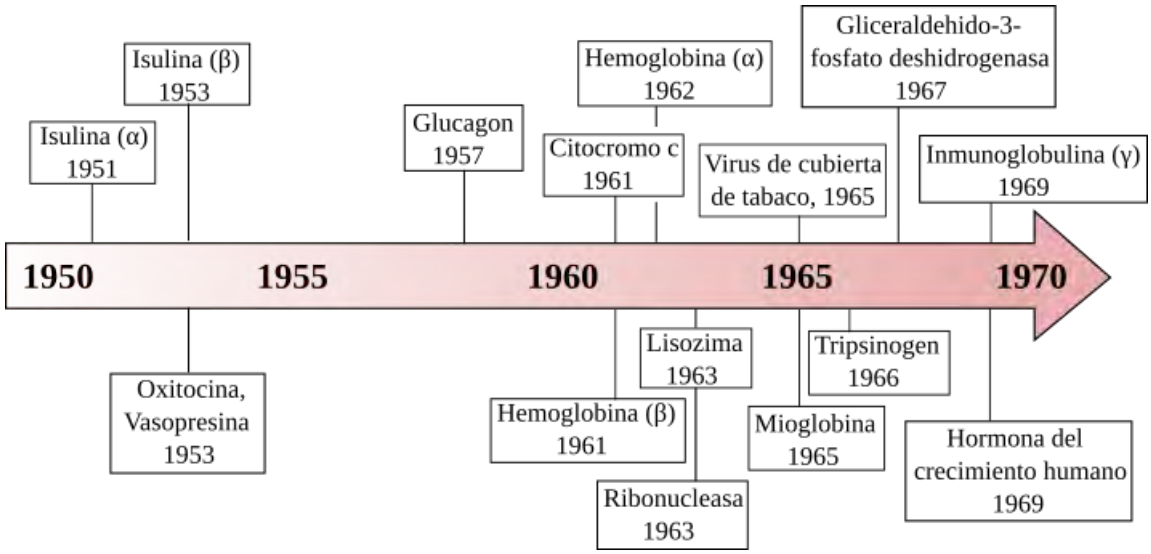



**Figura 1.1:** Primeras secuencias de proteínas. Modificada de Raymond *et al.*[9]

En 1964-1965, Margaret Dayhoff, considerada por muchos como la madre de la bioinformática, y sus colaboradores en la *National Biomedical Research Foundation* (NBRF), fueron los primeros en crear una base de datos de las secuencias de las proteínas conocidas en ese entonces, en un atlas de estructuras y secuencias que eventualmente se convirtió en el Expediente de Información de Proteínas (*Protein Information Resource*, PIR). En su primera edición, éste atlas contenía información de la secuencia de 65 proteínas pero para 1985 ya contaba con más de 2 500 secuencias de éstas[10][11]. Dayhoff, además, ideó el código de una letra para los aminoácidos para facilitar la manipulación de las secuencias, el cual se continúa usando. También comparó las secuencias conocidas de proteínas en búsqueda de relaciones existentes entre ellas con la finalidad de estudiar sus relaciones, lo que la llevó a crear una métrica de similitud conocida como matriz PAM (*Pointed accepted mutations*)[12][13].

En la actualidad, la base de datos del NCBI (*National Center for Bioinformatics Information*) es uno de los recursos web más usados en todo el mundo y cuenta ya con más de 190 millones de secuencias (Figura 1.2)[14].

La bioinformática se ha convertido en una de las áreas de investigación más importantes por el uso que hace de las matemáticas y la computación para facilitar tareas que sin estas herramientas tomaría mucho tiempo realizar. El continuo crecimiento en el poder de procesamiento de las computadoras ha permitido que la bioinformática tenga un papel cada vez más importante en las ciencias de la vida.[15]

## 1.1. Fundamentos bioquímicos

### 1.1.1. Proteínas

Las proteínas son de las moléculas más versátiles en los seres vivos, pues realizan funciones como el transporte y almacenamiento de compuestos, catálisis, crecimiento, protección inmunológica y formación de estructuras, entre otras[16][17][18]. Las proteínas son cadenas polipeptídicas constituidas por entre 50 y 2 000 **aminoácidos**, que son compuestos con estructuras muy similares entre ellos, lo que permite agruparlos en función de sus propiedades químicas u otras características[19]. Aunque hay una gran cantidad de aminoácidos, aquellos presentes en los organismos vivos son sólo veinte[20][21][22].

Estructuralmente, los aminoácidos están formados por cuatro componentes: un carbono central o **carbono** $\alpha$ (llamado así por ser el primer átomo de carbono de la molécula), un grupo **amino**, un grupo **carboxilo**, y una cadena lateral denotada como **grupo R** o sustituyente que diferencia a cada uno de los veinte aminoácidos y les brinda sus distintas propiedades químicas (Figura 1.3)[23][24].

Un grupo de aminoácidos unidos se llama *péptido* y al enlace que los une se le llama **enlace peptídico** o **enlace amida**, el cual es un enlace covalente entre el grupo amino de un aminoácido y el grupo carboxilo del otro en el que se libera

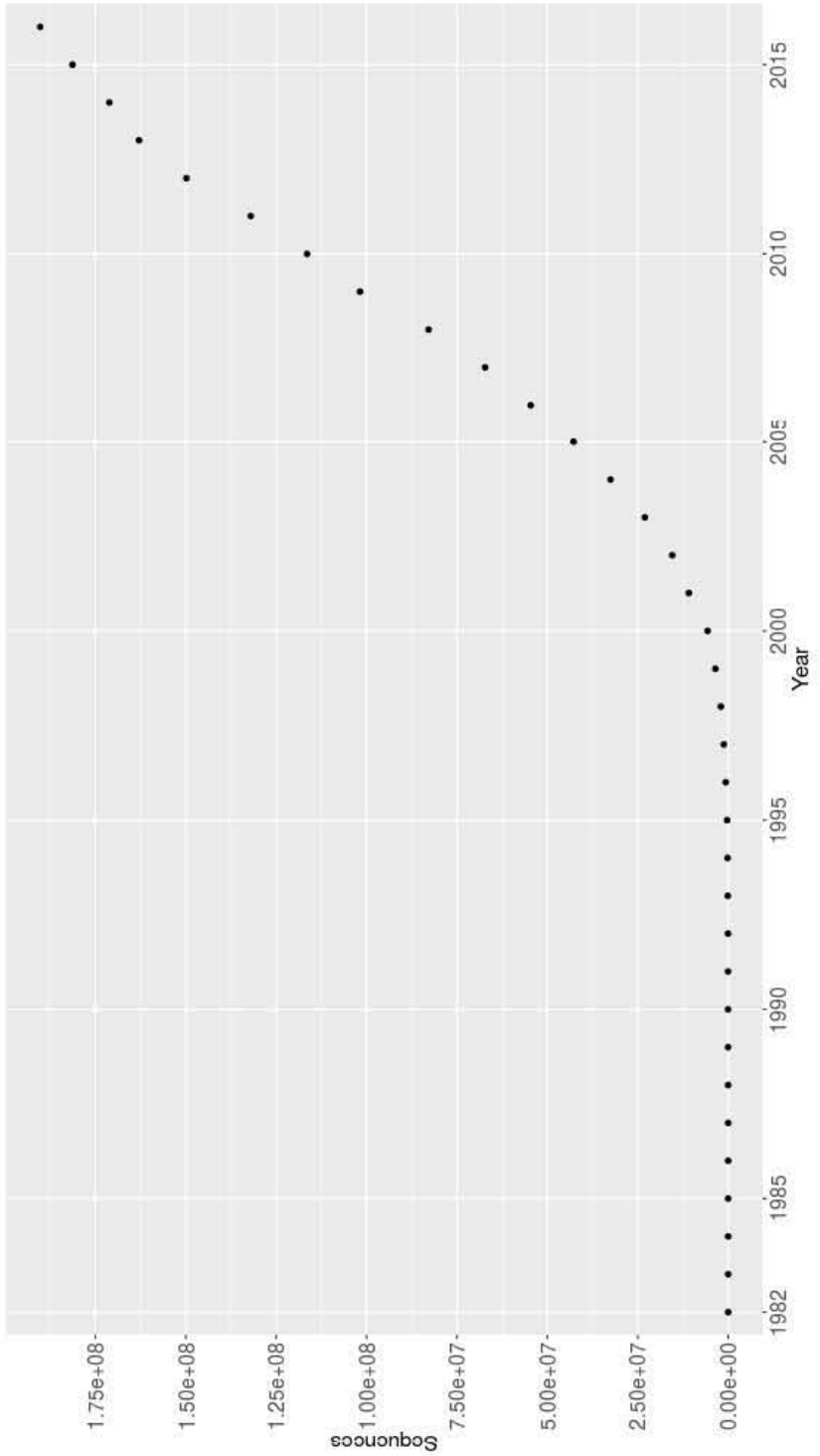



**Figura 1.2:** Crecimiento de la cantidad de secuencias alojadas en el GenBank del NCBI. Con información de `http://www.ncbi.nlm.nih.gov/genbank/statistics/`.


**Figura 1.3:** Clasificación de los aminoácidos de acuerdo a las propiedades de sus sustituyentes R. A la derecha se muestran los nombres y sus abreviaciones. Modificado de Esquivel *et al.*[25].

agua al medio, esa liberación es lo que hace que a cada aminoácido en una cadena peptídica se le llame también *residuo*.

Los enlaces peptídicos son enlaces sencillos, lo que permite una rotación libre de los péptidos adyacentes en ángulos llamados *ángulos de torsión*, que le brindan a las proteínas una amplia variedad de plegamientos.

#### 1.1.1.1. Estructura primaria

Una proteína en su forma de estudio más sencilla se llama **estructura primaria** y consiste en la serie de residuos o aminoácidos que la conforman linealmente. Debido a que ciertos aminoácidos tienen propiedades físico-químicas similares, las secuencias de una misma proteína pueden presentar ligeras variaciones en esta estructura. Esa propiedad de permitir residuos "intercambiables" es conocida como *polimorfismo*, y es la causa principal de variaciones en los humanos. Se estima que entre entre 20 %-30 % de las proteínas en humanos son polimórficas. A las proteínas que son muy similares en sus residuos se les agrupa en *familias*.

#### 1.1.1.2. Estructura secundaria

Al ordenamiento de los residuos de las proteínas en el espacio se le llama **conformación**, y a pesar de que los ángulos de torsión les permiten a las proteínas plegarse en una virtualmente infinita cantidad de formas, en 1963 G. Ramachandran demostró que sólo unas cuantas conformaciones están permitidas debido a colisiones entre moléculas y a que el plegamiento se da en estados que requieran una energía

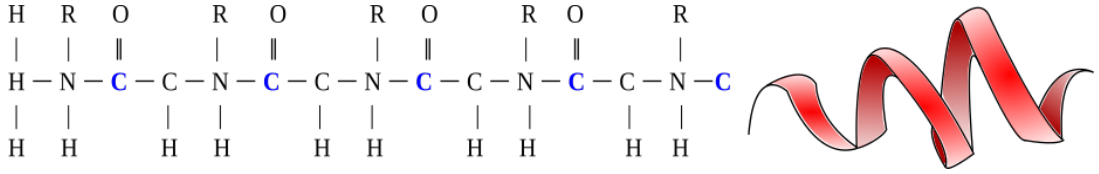


**Figura 1.4:** Representación plana y estrucural de la hélice $\alpha$.

mínima[26]. Por lo tanto, una proteína en su medio no se encuentra en una forma lineal sino que se pliegan por las interacciones con el medio y entre sus residuos.

El estudio de distintas proteínas demostró que había segmentos que se plegaban de una forma específica y que estaban presentes en distintas especies u organismos. A esos segmentos se les llamó **estructura secundaria**. Estas estructuras se estabilizan a través de puentes de hidrógeno entre los grupos amino y carboxilo de distintos residuos. Se han encontrado distintas estructuras secundarias en proteínas pero sólo dos de ellas se encuentran en una absoluta mayoría: la *hélice* $\alpha$ y la *hebra* $\beta$.

La **hélice** $\alpha$ es una estructura helicoidal en la que los grupos R están orientados hacia afuera de la hélice. Tanto las hélices dextrógiras como levógiras son posibles, aunque se presentan con mayor frecuencia las que tienen orientación dextrógira debido a que son menores las colisiones estéricas. Estas estructuras se representan como hélices o como cilindros (Figura 1.4).

La **hebra** $\beta$ o conformación $\beta$ es una estructura en la que los aminoácidos se extienden en zig-zag. Estas estructuras se enlazan con puentes de hidrógeno a otras hebras para formar **hojas** $\beta$ *paralelas* si las cadenas polipeptídicas proceden en la misma dirección amino-carboxilo, o *antiparalelas* si proceden en direcciones opuestas. En las hojas $\beta$, las cadenas laterales de los residuos sobresalen de forma paralela por ambos lados de la hoja. Las hojas $\beta$ se representan como una flecha (Figura 1.5).

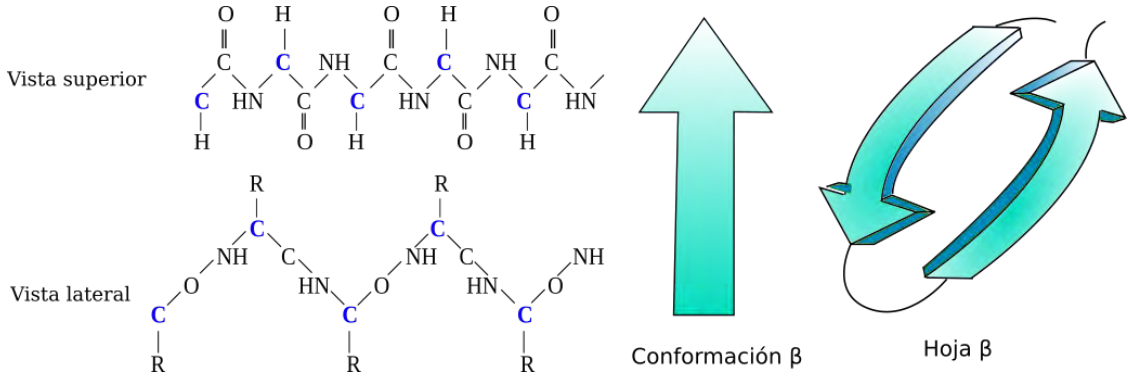



**Figura 1.5:** Representación plana y estructural de una hebra y una hoja $\beta$ antiparalela.

Otras estructuras secundarias presentes en las proteínas, aunque menos comunes, son el giro $\beta$, la vuelta $\Omega$ y la triple hélice.

A un grupo de estructuras secundarias dispuestas en una forma característica y a las interacciones entre ellas se les denomina **motivos**. Los motivos se pueden agrupar para formar **dominios**, que como se verá posteriormente, son partes de una cadena polipeptídica que poseen una función. Algunos de los motivos más comunes se muestran en la Figura 1.6.

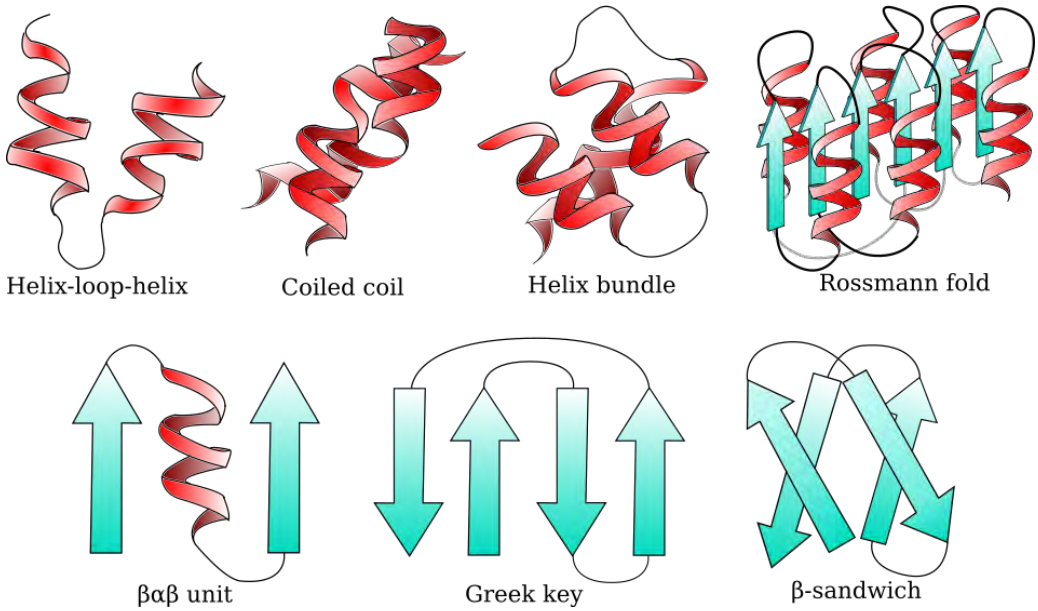



**Figura 1.6:** Motivos más comúnes en proteínas.

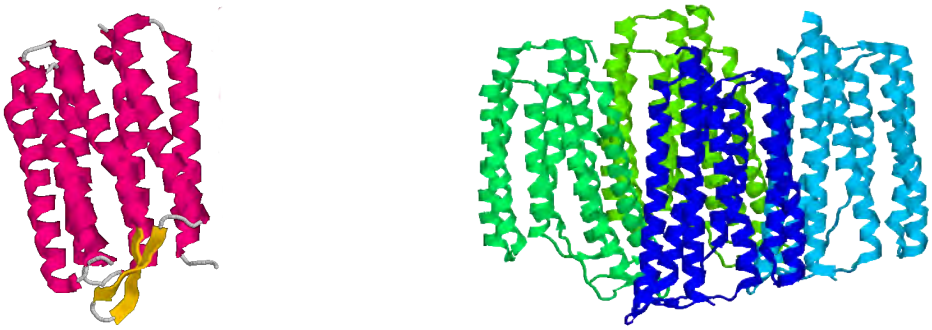
En un gran número de enzimas, los motivos comparten una súper estructura secundaria común compuesta por una serie de hebras $\beta$ y hélices $\alpha$. Esta estructura fue nombrada **Rossmann fold** después que Michael Rossmann la descubriera en 1974[27]. Los motivos Rossmann fold son una de conformaciones más profundamente estudiadas descritas como seis hebras $\beta$ con un orden relativo 3-2-1-4-5-6 rodeadas por seis hélices $\alpha$ (tres de cada lado)[28][29].

#### 1.1.1.3. Estructura terciaria y cuaternaria

Al arreglo tridimensional completo de todos los átomos en una proteína se le conoce como **estructura terciaria**.

Mientras la estructura secundaria se refiere a fragmentos de la secuencia de aminoácidos ordenada en una forma espacial, la estructura terciaria incluye además de la conformación espacial de la proteína entera y las interacciones químicas entre sus residuos. Una proteína plegada en su estructura terciaria se dice que está en su forma *nativa*. Cuando una proteína pierde su forma nativa se dice que se *desnaturaliza*, ésto sucede cuando el medio es muy ácido o alcalino, por cambios en la temperatura, o por factores químicos.

A la unión de varias cadenas polipeptídicas en un arreglo particular se conoce como **estructura cuaternaria** en la que cada cadena que la conforma recibe el nombre de *subunidad*, y al conjunto de proteínas que se unen en una estructura cuaternaria se le llama *complejo multiprotéico* (Figura 1.7).

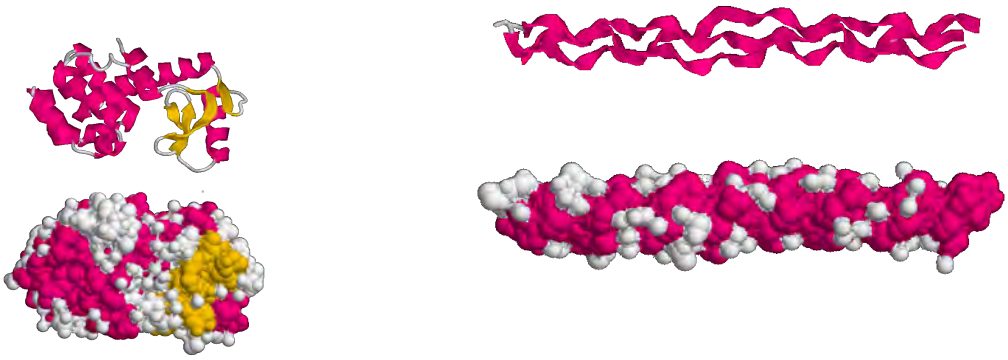


**(a)** Estructura de una de las cuatro cadenas de la bacteriorodopsina.

**(b)** Estructura cuaternaria de la bacteriorodopsina.

**Figura 1.7:** La bacteriorodopsina es una membrana protéica en bacterias que actúa como bomba de protones. Su conformación es esencial para su función. Imagen generada con RasMol

La observación de estos niveles superiores muestra que éstas se pueden agrupar en dos clases, proteínas fibrosas, con cadenas ordenadas en largas hojas o hebras; y proteínas globulares, con cadenas plegadas en una forma esférica o globular. Mientras las primeras tienen funciones más bien estructurales, las segundas poseen funciones químicas (Figura 1.8).



**(a)** Lisozima

**(b)** Colágeno

**Figura 1.8:** La lisozima es una enzima presente en la saliva, lágrimas y sudor que brinda protección contra infecciones bacterianas. El colágeno es el componente principal de huesos, dientes y cartílago, y es una de las proteínas fibrosas más abundantes. Imagen generada con RasMol.

Adicionalmente, el conocer las estructuras de las proteínas en sus niveles más elevados permite agruparlas en *súperfamilias*, que comparten relaciones estructurales similares.

### 1.1.2. Enzimas

El éxito de un organismo depende principalmente de dos factores: primero, debe ser capaz de autoreplicarse; segundo, debe ser capaz de catalizar reacciones de forma efectiva y selectiva.

Las **enzimas** son proteínas sintetizadas en las células que aceleran de forma selectiva y eficiente las reacciones que se dan en su entorno celular. La mayor parte de ellas son proteínas con estructuras globulares y aproximadamente una cuarta parte de los genes en el genoma humano codifican enzimas, lo que muestra su importancia en la vida.

Una reacción química catalizada por una enzima toma lugar en una parte de la enzima llamada **sitio activo**, cuya conformación es altamente específica para una o un grupo de enzimas. La catálisis convierte uno o más compuestos llamados **sustratos** en uno o más compuestos diferentes llamados **productos** sin consumir a la propia enzima.

Existen otras moléculas de naturaleza no proteica que también participan en la catálisis, estas moléculas son iones metálicos llamados *cofactores* que se unen a la enzima para acelerar el proceso catalítico. El aumento de la temperatura, cambios en el pH y concentración de enzima-sustrato pueden alterar la velocidad de catálisis ralentizándola o acelerándola, aunque en este último caso el límite lo establecen las condiciones de desnaturalización.

La *Enzyme Commission* (EC), de la Unión Internacional de Bioquímica (IUB, por sus siglas en inglés), creó una clasificación de las enzimas de acuerdo al tipo de reacción que catalizan. El Cuadro 1.1 muestra las seis clases de enzimas. La clasificación de la IUB consiste del prefijo *EC* seguido de cuatro dígitos que representan la clase a la que pertenece, el grupo funcional en el que actúan, el receptor, y el sustrato. Debido a que la clasificación de la IUB presenta nombres largos y en ocasiones de difícil uso, las enzimas comúnmente se nombran de la forma tradicional (aunque ambigua), agregando el sufijo "asa" al sustrato o tipo de reacción que catalizan.

La clasificación de la IUB para las enzimas se puede consultar en el sitio `http://www.chem.qmul.ac.uk/iubmb/enzyme/`.

#### 1.1.2.1. Mecanismo de catálisis de las enzimas

Aunque las enzimas poseen diferentes funciones, el sitio activo comparte características comunes a todas ellas: (1) tiene una forma de hendidura para permitir la fijación del sustrato; (2) sin importar el tamaño de la enzima, representa sólo una pequeña parte de ella, los demás aminoácidos sirven como estructura, vías de transporte, o sitios de interacción con otras proteínas; (3) la unión del sustrato al sitio activo se realiza mediante fuerzas no covalentes.

| N$^o$ | Clase | Descripción |
|---|---|---|
| 1 | Oxido-rreductasas | Catalizan reacciones de oxidación y reducción: Deshidrogenasas, oxidasas, reductasas, peroxidasas, catalasa, oxigenasas, hidroxilasas. |
| 2 | Transferasas | Transfieren un grupo químico de una molécula a otra: Transaldolasas y transcetolasas, fosforiltransferasas, quinasas, fosfomutasas. |
| 3 | Hidrolasas | Rompen enlaces a través de hidrólisis: Esterasas, glucosidasas, peptidasas, fosfatasas, tiolasas, fosfolipasas, amidasas, desaminasas, ribonucleasas. |
| 4 | Liasas | Agregan o remueven grupos para formar dobles enlaces: Descarboxilasas, aldolasas, hidratasas, deshidratasas, sintasas, liasas. |
| 5 | Isomerasas | Provocan cambios estructurales (isómeros): Racemasas, epimerasas, isomerasas, mutasas. |
| 6 | Ligasas | Forman nuevos enlaces hidrolizando ATP: Sistetasas, carboxilasas. |

**Cuadro 1.1:** Clasificación de las enzimas según la reacción que catalizan.

Daniel Koshland (1958) propuso el modelo de *adaptación inducida*, que atribuía los cambios conformacionales del sustrato a su interacción con un sitio activo flexible (Figura 1.9).

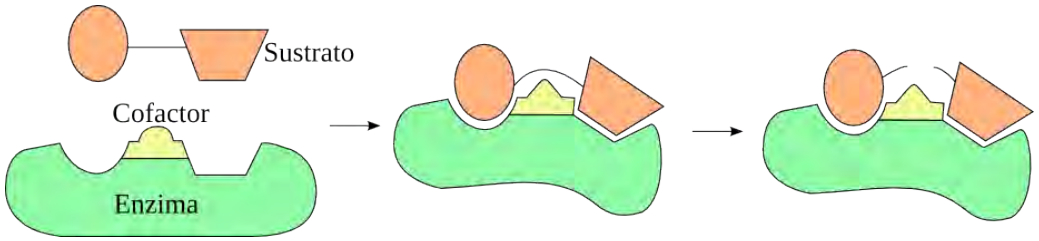



**Figura 1.9:** Representación del modelo de Koshland de adaptación inducida. Modificado de Murray *et al.*[23].

A pesar de la especificidad de las enzimas, su actividad catalítica puede ser inhibida por pequeñas moléculas o por iones llamados **inhibidores**, que ralentizan o detienen las reacciones enzimáticas. El inhibidor es *irreversible* si su unión a la enzima se da mediante enlaces covalentes o no-covalentes muy estables a residuos imprescindibles para la catalización, o *reversible* si se une a la enzima de forma no-covalente y semi-estable.

### 1.1.3. Dominios funcionales

El concepto de domino en sus orígenes se usaba para describir regiones en la estructura terciaria de una proteína, posteriormente se descrubrió que algunas de esas regiones estaban presentes en distintas estructuras o incluso repetidos en la misma cadena polipeptídica[30].

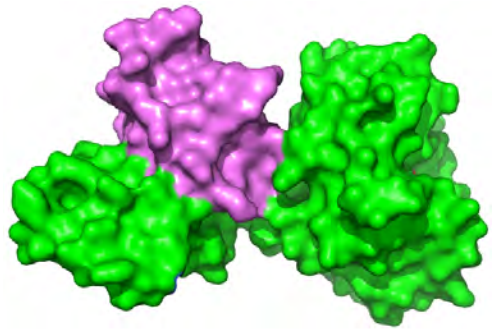


**Figura 1.10:** Proteína CTP1L. En verde la proteína, en violeta un dominio. EMBL/Rob Meijers `http://news.embl.de/science/1601-endolysin/`.

Actualmente, un **dominio** se define como un fragmento de una secuencia proteica que puede plegarse y adquirir su función aún en aislamiento del resto de la proteína[30][31]. Esta característica hace que a los dominios se les considere como los bloques de construcción de las proteínas (Figura 1.10).

La mayoría de métodos existentes para la asignación de dominios se basan en medidas geométricas de las estructuras terciarias y cuaternarias de las proteínas. La comparación de sus estructuras ha permitido observar la recurrencia de dominios en distintas proteínas o identificar proteínas multi-dominio que se han preservado y duplicado para generar combinaciones de orden superior[32] por especiación o por mutaciones y duplicaciones[31]. El estudio de cómo la duplicación y el reordenamiento ha ocurrido permite entender las relaciones funcionales entre dominios y determinar cómo las proteínas han evolucionado (Figura 1.11).

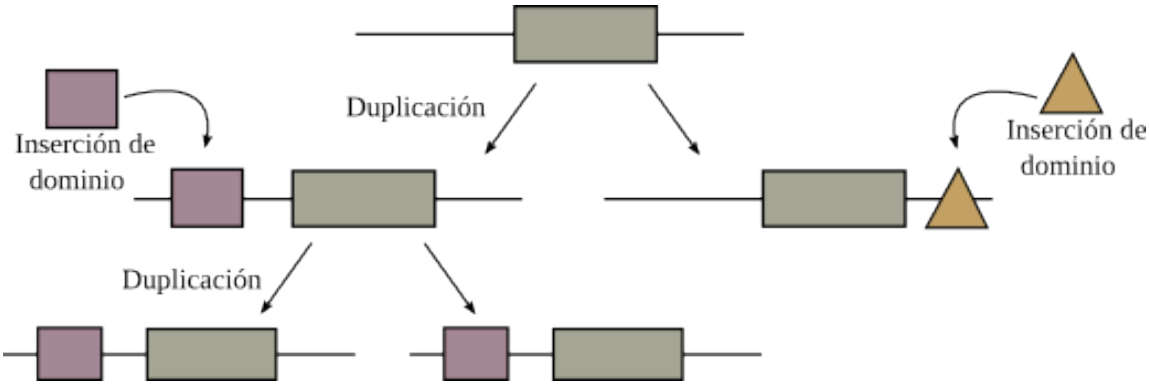



**Figura 1.11:** Árbol evolutivo hipotético de una proteína multidominio que evolucionó a través de duplicación e inserción de dominios. Modificado de Song *et al.* [31].

La presencia de ciertos dominios distribuidos en los tres dominios de la vida sugiere una ancestría remota que se ha conservado debido a que pueden evolucionar con facilidad, a que están adaptados para muchos nichos funcionales o a que son fundamentales para los procesos celulares. La existencia de dominios funcionales indica que han surgido durante la evolución por la combinación de genes que codificaban diferentes proteínas ancestrales y que es posible que alguna vez fueron una molécula separada[33]. Además, se ha encontrado que proteínas con los mismos dominios pero "barajados" por procesos evolutivos presentan diferentes propiedades funcionales[34][35].

Uno de los resultados del genoma humano, fue el descubrimiento de que muchas

proteínas multidominio están involucradas en procesos como la respuesta inmune y la reparación de tejidos[5].

## 1.2. Fundamentos matemáticos y computacionales

La clasificación de proteínas en familias y súperfamilias brinda información sobre su estructura, actividad y rol metabólico[36]. Actualmente hay una gran cantidad de recursos web para estudiar proteínas en sus distintos niveles estructurales, sin embargo, la base para casi todos éstos métodos es el análisis de secuencias. El análisis de secuencias es la parte más fundamental de la bioinformática. El proceso más usado en este tipo de análisis es el alineamiento de un par de secuencias.

### 1.2.1. Alineamiento de secuencias

Si se consideran los 20 aminoácidos involucrados en los procesos biológicos y se realiza una comparación de dos secuencias creadas al azar, la probabilidad de que sean similares es de apenas 5 %[37]. Así, el alineamiento de secuencias provee información para inferir el parentesco de las secuencias en estudio de tal forma que si dos secuencias comparten una similitud significativa es improbable que se deba al azar.

El alineamiento de secuencias es lo que permite agrupar a las proteínas en familias si comparten al menos un 25 % de similitud en sus residuos. Los miembros de estas familias de proteínas son llamados proteínas homólogas o simplemente **homólogos**.

Un **alineamiento de dos secuencias** está formado por la inserción de espacios o guiones (*gaps*) en locaciones arbitrarias a lo largo de la longitud de la secuencias hasta que éstas tengan la misma longitud y no haya dos espacios en la misma posición en las secuencias aumentadas[38][39]. Por ejemplo, si $S = matematicas$ y $T = aplicadas$, un posible alineamiento sería:

```
matematicas
-ap-licadas
```

La relación entre dos secuencias se describe con tres conceptos:

**Similitud.** Es el porcentaje de residuos con propiedades químicas idénticas o semejantes en el alineamiento.

**Identidad.** Es el porcentaje de residuos idénticos en el alineamiento.

**Homología.** Es una propiedad que se basa en la similitud para suponer si dos secuencias provienen de un ancestro común. Si la similitud es mayor a 30 %,

la evidencia de que las secuencias sean homólogas es fuerte; si está entre 20 % y 30 %, la homología no es clara; por último, si es menor al 20 %, no se puede asegurar que las secuencias sean homólogas.

El alineamiento de pares de secuencias permite identificar residuos altamente conservados en las secuencias, así como identificar mutaciones (sustituciones, inserciones o deleciones).

El puntaje de los alineamientos por residuo suele hacerse con matrices creadas a partir de observaciones de frecuecias de aminoácidos en distintas familias de proteínas (matrices PAM o BLOSUM).

Existen dos algoritmos de alineamiento para un par de secuencias, el alineamiento global y el alineamiento local.

#### 1.2.1.1. Alineamiento global (Needleman-Wunsch)

En 1970, Saul Needleman y Christian Wunsch publicaron un algoritmo para producir un alineamiento óptimo entre un par de secuencias de proteínas o nucleótidos. El algoritmo de Needleman-Wunsch se usa para alinear dos secuencias de longitudes iguales o muy similares.

Este algoritmo trabaja con una matriz en la que se busca alinear ambas las secuencias colocando un cero en la esquina superior izquierda y en el resto de la primer fila y columna, valores descendiendes que dependen de una penalización por inserción de espacios. Los valores de las demás celdas se calculan con las ecuaciones de recurrencia:

$$s_{i,j} = max \begin{cases} s_{i-1,j} + \delta(v_i, -) \\ s_{ij-1} + \delta(-, w_i) \\ s_{i-1,j-1} + \delta(v_i, w_j) \end{cases} \tag{1.1}$$

donde $\delta$ es el costo de alineamiento del elemento $i$ con el elemento $j$.

Se obtiene el camino óptimo haciendo uso de programación dinámica y *backtracking*. Un paso en diagonal significa una identidad o sustitución, un paso hacia la derecha o hacia arriba es la inserción de un espacio.

Como el algoritmo de Needleman-Wunsch considera la longitud total de las secuencias a alinear, es conocido también como algoritmo de alineamiento global. Este algoritmo es óptimo cuando las dos secuencias a alinear son generalmente similares en residuos y en tamaño. El algoritmo alinea ambas secuencias de inicio a fin y conserva el alineamiento con el mayor puntaje de similitud.

#### 1.2.1.2. Alineamiento local (Smith Waterman)

El algoritmo de alineamiento local o algoritmo de Smith-Waterman fue desarrollado en 1981 por Temple Smith y Michael Waterman. A diferencia del algoritmo de Needleman-Wunsch, el algoritmo de Smith-Waterman se enfoca en alinear regiones cortas de las secuencias. Ya que el algoritmo no busca alinear dos secuencias a lo largo de sus longitudes, es muy útil para buscar regiones altamente conservadas en proteínas y para alinear secuencias con longitudes distintas.

De la misma forma que el algoritmo de Needleman-Wunsch, el algoritmo de alineamiento local comienza construyendo una matriz en la que su primer fila y columna tiene valores "0". Los valores para las demás celdas se calculan con las ecuaciones de recurrencia:

$$s_{i,j} = max \begin{cases} s_{i-1,j} + \delta(v_i, -) \\ s_{ij-1} + \delta(-, w_i) \\ s_{i-1,j-1} + \delta(v_i, w_j) \\ \quad 0 \qquad\qquad \text{si } s_{i,j} < 0 \end{cases} \tag{1.2}$$

que son idénticas a las ecuaciones de recurrencia del alineamiento global salvo por la condición extra que no permite que la matriz posea valores negativos.

Para encontrar alineamientos, se identifica el valor más alto en la matriz y se retrocede en diagonal hasta encontrar un "0", lo que indica el fin del alineamiento, luego se busca el siguiente valor más alto y se repite el proceso. El resultado será uno o más fragmentos alineados que no necesariamente tienen la longitud total de las secuencias.

#### 1.2.1.3. Alineamientos múltiples

El alineamiento de múltiples secuencias es una extensión del alineamiento de secuencias que consiste en alinear pares de secuencias de forma recursiva. Las ventajas del alineamiento de múltiples secuencias es que permite la identificación de regiones conservadas y motivos que no serían evidentes al comparar sólo dos secuencias. El alineamiento múltiple es esencial en el análisis filogenético de familias de proteínas y en la predicción de estructuras secundarias y terciarias.

En un alineamiento de múltiples secuencias, los residuos alineados presentan dos homologías, una en un sentido evolutivo, siendo posible que provengan de un ancestro común, y otra en un sentido estructural, pues los residuos alineados tienden a ocupar las mismas posiciones en la estructura tridimensional de cada proteína alineada[40].

Existen distintos software de alineamiento de múltiples secuencias tales como CLUSTAL, MUSCLE o T-COFFEE, entre otros[41][42].

| E-value | Relación consulta-BD |
|---|---|
| $E < 1e-50$ | Alta confianza en una relación homóloga. |
| $1e-50 < E < 0{,}01$ | Pueden considerarse como homólogos. |
| $0{,}01 < E < 10$ | Relación no significativa, quizá homólogos remotos. |
| $E > 10$ | No hay relación o su relación es muy remota. |

**Cuadro 1.2:** Interpretación común del *e-value* entre la secuencia de consulta y la(s) secuencia(s) en la base de datos.

### 1.2.2. Herramientas de alineamiento para búsquedas en bases de datos: FASTA y BLAST

El algoritmo de Smith-Waterman garantiza encontrar el alineamiento óptimo de dos o más secuencias, sin embargo, al ser un algoritmo que hace uso de programación dinámica es computacionalmente complejo[43].

Este hecho propició el desarrollo de dos algoritmos que hacen uso del método de Smith-Waterman de y una heurística que les permite reducir el tiempo de cómputo, aunque por lo mismo no se garantiza que el resultado sea el mejor alineamiento posible. Ambos algoritmos son usados para buscar similitudes entre una secuencia de consulta y una base de datos.

#### 1.2.2.1. FASTA

FASTA (*FAST All*) fue desarrollado en 1988 por Pearson & Lipman[44] y es uno de los algoritmos más usados para realizar alineamientos de secuencias.

Para medir la significancia estadística del alineamiento, FASTA usa un estadístico llamado **e-value**, el cual indica la probabilidad de que los resultados de una búsqueda sean debidos al azar. Es decir, que mientras menor sea el valor del *e-value*, menor es la probabilidad de que el alineamiento sea azaroso (Cuadro 1.2).

Debido a que si el tamaño de la base de datos aumenta aumentará también el *e-value*, en algún momento podrían perderse relaciones homólogas[44], por tal motivo FASTA incluye el **bit score**, que es un indicador que mide la similitud sin considerar la longitud de la secuencia de consulta ni el tamaño de la base de datos. Así, mientras mayor sea el valor del *bit score*, mayor será la probabilidad de que una serie de proteínas sean homólogas.

##### Formato FASTA

FASTA también posee un tipo de archivo para secuencias de proteínas o nucléotidos. El formato consiste de una línea que inicia con el símbolo '>' seguido de la

descripción y la secuencia sin espacios y en líneas con menos de 80 caracteres. Un ejemplo de una secuencia en formato FASTA es el siguiente:

```
>gi|821521776|ref|WP_046843704.1|2'-5'RNA ligase ['Deinococcus soli']
MRVKYPSTPHLPWSPGLQNDDRRIPSLRDLEGQEVVVTEKLDGENTSLYRADLHARSLDTRPHPSRTWVK
AERGRFGHEIPPGWRLCGENVFAVHSLRYEALAGYFYLFSVWDDRNVSRPWAEVRDWAARLDVPTPRELY
RGVWDEAALRALTVDTARMEGYVVRVTGEIPYAQFGRRVAKWVRAGHVQTDEHWLSRPVERNGLRDEPA
```

La importancia del formato FASTA es que se ha vuelto un estándar para almacenar secuencias ya sea de aminoácidos o de nucleótidos. FASTA está disponible en `http://www.ebi.ac.uk/Tools/sss/fasta/`.

#### 1.2.2.2. BLAST

BLAST (*Basic Local Alignment Search Tool*) es el algoritmo de alineamiento del Centro Nacional para la Información Biotecnológica (NCBI, por sus siglás en inglés) para comparar secuencias. Fue desarrollado por Stephen Altschul *et al.* en 1990[45].

BLAST considera sustiticiones con puntajes positivos (mutaciones aceptadas) para obtener todas las secuencias similares a la palabra de consulta que están en la base de datos. Posteriormente calcula los puntajes de cada una usando una matriz y conserva la palabra con mayor puntaje para que a partir de ella se comparen las regiones anteriores y posteriores a la palabra de consulta.

Al igual que FASTA, BLAST mide la significancia estadística con el *e-value* y el *bit score*. BLAST hace búsquedas en distintas bases de datos tales como GenBank, Protein Data Bank, SwissProt y Protein Information Resource, entre otras. BLAST está disponible en el sitio web del NCBI `https://blast.ncbi.nlm.nih.gov/Blast.cgi`.

### 1.2.3. Modelos para la identificación de dominios funcionales

Una secuencia en la que se buscan dominios se puede comparar a un documento en el que se buscan palabras. En un documento las palabras no son igualmente importantes, además de que hay algunas palabras que proveen más información sobre el contexto del documento que otras. La frecuencia de una palabra en un documento también es una indicación de su importancia en ese documento.

Una de las aplicaciones de los alineamientos de múltiples secuencias es para la creación de Matrices de puntaje de posición específica (PSSM, *Position-Specific Score Matrices*), perfiles, y modelos ocultos de Markov (HMM, *Hidden Markov Models*). Estos son modelos estadísticos que reflejan la información de frecuencia de nucleótidos o aminoácidos en alineamientos múltiples.

| Posición | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Seq. 1 | A | T | G | T | C | G |
| Seq. 2 | A | A | G | A | C | T |
| Seq. 3 | T | A | C | T | C | A |
| Seq. 4 | C | G | G | A | G | G |
| Seq. 5 | A | A | C | C | T | G |

**a)** Se genera un alineamiento múltiple

| Pos. | 1 | 2 | 3 | 4 | 5 | 6 | Freq. |
|---|---|---|---|---|---|---|---|
| A | 0.6 | 0.6 | – | 0.4 | – | 0.2 | 0.30 |
| T | 0.2 | 0.2 | – | 0.4 | 0.2 | 0.2 | 0.20 |
| G | – | 0.2 | 0.6 | – | 0.2 | 0.6 | 0.27 |
| C | 0.2 | – | 0.4 | 0.2 | 0.6 | – | 0.23 |

**b)** El alineamiento múltiple se convierte en una tabla de frecuencias

| Pos. | 1 | 2 | 3 | 4 | 5 | 6 | Freq. |
|---|---|---|---|---|---|---|---|
| A | 2.0 | 2.0 | – | 1.33 | – | 0.67 | 0.30 |
| T | 1.0 | 1.0 | – | 2.0 | 1.0 | 1.0 | 0.20 |
| G | – | 0.74 | 2.22 | – | 0.74 | 2.22 | 0.27 |
| C | 0.87 | – | 1.74 | 0.87 | 2.61 | – | 0.23 |

**c)** Se normalizan los valores dividiéndolos entre su frecuencia

| Pos. | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| A | 1.0 | 1.0 | – | 0.41 | – | -0.58 |
| T | 0.0 | 0.0 | – | 1.0 | 0.0 | 0.0 |
| G | – | -0.43 | 1.15 | – | -0.43 | 1.15 |
| C | -0.2 | – | 0.8 | -0.2 | 1.38 | – |

**d)** Se aplica logaritmo base 2

**Figura 1.12:** Ejemplo de la construcción de una PSSM a partir de un alineamiento de múltiples secuencias. Modificado de Xiong[37].

#### 1.2.3.1. Matrices de puntaje de posición específica y Perfiles

Una PSSM es una matriz que contiene información probabilística de los aminoácidos o nucleótidos en cada posición en un alineamiento de múltiples secuencias sin espacios. En esta matriz, las filas representan las posiciones de los residuos de un alineamiento múltiple y las columnas representan los nombres de los residuos, o viceversa. Los valores en la matriz son puntajes de log-probabilidades de los residuos (Figura 1.12).

Este modelo puede ser usado para búsqueda en bases de datos o para verificar cómo se ajusta una nueva secuencia a una familia. Los valores de la PSSM depende del número de secuencias usadas para generar la matriz.

Cuando en un alineamiento de múltiples secuencias se incluye información sobre la penalización por espacios, se crea un **perfil**. En otras palabras, un perfil es una PSSM con información sobre la penalización por inserción o supresión de residuos para una familia de secuencias. Los perfiles también pueden ser usados en una base de datos para buscar secuencias homólogas remotas.

Para la puntuación de los perfiles es común el uso de matrices BLOSUM[46], PAM y algunas otras similares[47][48].

#### 1.2.3.2. Modelos ocultos de Markov

Una forma más eficiente de calcular los puntajes de similitudes entre una secuencia y el perfil de una secuencia es a través del uso de modelos ocultos de Markov (HMM). Estos modelos fueron diseñados originalmente para ser usados en problemas de detección de señales que iban desde el reconocimiento de voz hasta sonares[49][50][51]. En la bioinformática, los modelos ocultos de Markov se usan para alineamientos de secuencias, predicción de estructuras y dominios en proteínas, y análisis en cromosomas y genes[52].

El modelo oculto de Markov se basa en las cadenas de Markov, pero a diferencia de éste, el HMM combina una sola cadena de estados observables con dos o más cadenas de estados no observables (u ocultos) que influencian el resultado de los estados observables (Figura 1.13).

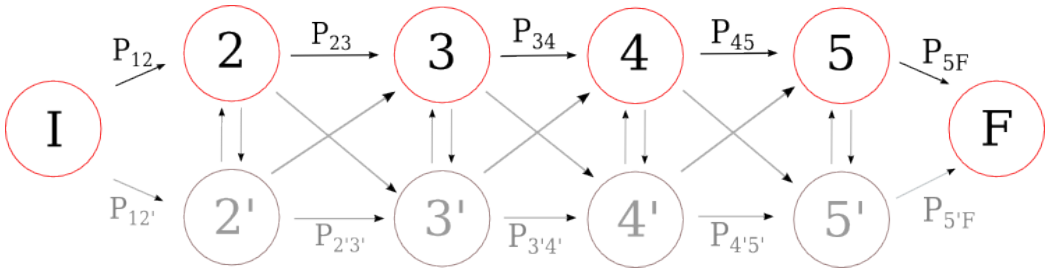



**Figura 1.13:** Representación simple de un modelo de Markov con una cadena de estados observables (rojo) y una de estados no obsevables (gris).

Los HMM también pueden ser usados para comparar secuencias, en particular para alinear una secuencia contra un perfil[53].

### 1.2.4. HMMER

HMMER (`http://hmmer.org/`) es un conjunto de herramientas para análisis de secuencias que hace uso de modelos ocultos de Markov. HMMER es usado para buscar secuencias homólogas de proteínas o nucleótidos y hacer alineamientos de secuencias.

Comparado con BLAST, FASTA, y otras herramientas de búsqueda y alineamiento de secuencias, HMMER ha demostrado ser más preciso y capaz de detectar homólogos remotos debido a la fuerza de sus modelos probabilísticos[54]. HMMER puede hacer uso de bases de datos de modelos de Markov −como Pfam− para identificar dominios y está disponible para sistemas UNIX/Linux y OS.

### 1.2.5. Recursos para análisis de dominios funcionales

Existen distintos recursos para analizar los dominios funcionales, las cuales se diferencian por la forma de búsqueda y los resultados obtenidos. Algunas de las más consultadas son:

**PFAM** Hace uso de HMM para alineamiento y búsqueda de secuencias. Ofrece una amplia varidedad de resultados como árboles filogenéticos, familias relacionadas, clanes, perfiles, especies, entre otras. La versión actual contiene más de 16 000 entradas[55][56], además, Pfam permite la descarga de los HMM para la identificación de dominios en secuencias. URL: `http://pfam.xfam.org/`

**PROSITE** Consiste de documentos que describen dominios de proteínas, familias y sitios funcionales asociados a patrones y perfiles para identificarlos. A la fecha PROSITE cuenta con 1 759 entradas, 1 309 patrones, 1 157 perfiles[57]. URL: `http://prosite.expasy.org/`

**SMART** (*The Simple Modular Architecture Research Tool*). Es un sitio que provee identificación y una extensa anotación de dominios de proteínas y sus arquitecturas. SMART contiene más de 1 200 dominios curados manualmente. Hace uso de las bases de datos UniProt, Ensembl y STRING para lograr cien millones de dominios y características de proteínas anotadas[58]. URL: `http://smart.embl-heidelberg.de/`

**CDD** (*Conserved Domain Database*). Es una herramienta del NCBI para la anotación de unidades funcionales en proteínas. Consiste en una colección de modelos de dominios que incluye un conjunto de ellos curados manualmente, y que utiliza estructuras 3D para proveer una visión a las relaciones secuencia-estructura-función[59]. URL: `http://www.ncbi.nlm.nih.gov/cdd/`

**InterPro** Provee un análisis funcional de proteínas, clasificándolas en familias y prediciendo dominios en sitios importantes; también incluye información sobre genes. InterPro (v 48.0) contiene 36 766 miembros integrados en 26 238 entradas[60]. URL: `https://www.ebi.ac.uk/interpro/`

**SCOP.** (*Structural Classification of Proteins*). Es una base de datos de estructuras y relaciones evolutivas conocidas entre proteínas con estructura conocida. La versión más reciente (1.73) contiene 92 927 dominios organizados en 3 464 familias , 1 777 súperfamilias y 1 086 formas de plegado[61]. URL: `https://scop.berkeley.edu/`

# Capítulo 2

# Introducción

El estudio de las proteínas es un campo en el que se han estado involucrando cada vez más la creación de modelos probabilísticos para predicción y el poder de procesamiento de las computadoras. Estos estudios se enfocan principalmente en las secuencias y estructuras de las proteínas con la finalidad de identificar regiones conservadas, predecir funciones, reconocer variantes, y encontrar patrones de expresión génica.

Geer *et al.* (2002)[62] desarrollaron CDART (*The Conserved Domain Architecture Retrieval Tool*), que se enfoca en la comparación a nivel arquitectura. Con CDART es posible encontrar proteínas similares a la proteína query comparando los dominios que tiene. CDART hace uso de PSSM con BLAST para definir los dominios y usa la base de datos CDD para las anotaciones.

CDART recibe los identificadores de las secuencias y genera las arquitecturas de dominios de forma gráfica de la secuencia de consulta y secuencias similares con un puntaje definido como la cantidad de coincidencias entre la secuencia de consulta y las encontradas en la base de datos. CDART puede consultarse en `http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi`.

Al igual que Geer, Lee *et al.* (2009)[63] describe un método para comparar proteínas a nivel de arquitectura e introduce una métrica que llama WDAC (*Weighed Domain Architecture Comparison*), la cual mide la Frecuencia Inversa de Abundancia (IAF, por sus siglas en inglés) para identificar homólogos. WDAC mide también la presencia de un dominio en distintas familias.

WDAC trabaja con archivos en formato FASTA o con extensión `.seq` o `.txt`. WDAC limita la cantidad de posibles *e-values* a elegir y requiere que se conozca el dominio biológico al que pertenecen las secuencias.

WDAC hace uso de HMM y la base de datos Pfam, y puede ser consultado en `http://wdac.kr/`.

Messih *et al.* (2012)[64] plantearon la hipótesis de que las funciones de las proteínas no sólo están determinadas por la existencia de ciertos dominios, sino también por la recurrencia y orden de esos dominios. Usando un modelo de probabilidades *a posteriori* (DRDO) y uno de Naïve-Bayes (DRDO-NB), desarrollaron un software ejecutable que hace uso de la base de datos swissPfam para identificar los dominios. En sus resultados, mencionan que obtuvieron una mejor predicción de funciones, demostrando así que la recurrencia y orden de los dominios provee importante información sobre las funciones de las proteínas.

DRDO requiere que las secuencias de proteínas se obtengan de UniProt y posteriormente realiza una serie de búsquedas en la base de datos. El software puede obtenerse en `https://sfb.kaust.edu.sa/Pages/Software.aspx`.

La consideración de factores adicionales a la similitud entre proteínas ha conducido a la realización de análisis más exhaustivos para determinar homologías. Terrapon *et al.* (2014)[65] realizaron alineamientos de dominios para la identificación de homólogos con dos herramientas, RADS (*Rapid Alignment of Domain Strings*) y RAMPAGE (*Rapid Alignment Method of Proteins based on domain ArranGEments*).

RADS-RAMPAGE hace uso de la base de datos UniProt y en sus resultados, mencionan que obtuvieron casi la misma sensibilidad que BLAST pero en menor tiempo de cómputo.

RADS-RAMPAGE está disponible en `http://rads.uni-muenster.de/`.

A pesar de que los métodos para identifiación y predicción de homólogos en proteínas han sido muy útiles hasta la fecha, Lee & Lee[63] resaltan que los métodos actuales suponen que secuencias significativamente similares son homólogos, pero además existen otros factores que son una desventaja:

- Las búsquedas semánticas por la función que realiza no muestra todos los dominios involucrados sino sólo aquel con el que se hizo la anotación de la proteína.
- Son herramientas web, lo que limita al usuario a la disponibilidad del sitio así como a la velocidad de su conexión a Internet.
- Debido a la demanda de estos recursos, en ocasiones las búsquedas implican una gran cantidad de tiempo o no es posible realizarlas por la concurrencia de usuarios. Una búsqueda puede tomar algunos segundos o incluso horas.
- El manejo de herramientas locales (como en el caso de DRDO y RADS) carecen de documentación en la que se explique su uso por lo que a un bioinformático sin experiencia le resultaría bastante complicado.

- Estos métodos hacen búsquedas de dominios en bases de datos para obtener resultados, pero en ocasiones el usuario cuenta con bases de datos locales en las que desea buscar arquitecturas de dominios, lo que requiere que el usuario cuente con conocimientos de programación.

El creciente empeño en el estudio de los dominios funcionales más allá de sus componentes residuales aunado a las desventajas de los métodos actuales permite el desarrollo de herramientas que se enfocan en un estudio más detallado de los dominios funcionales. Sin embargo, esas herramientas cuentan con limitantes que restringen la identificación y asignación de dominios funcionales a aquellos que están más conservados, ignorando posibles relaciones entre proteínas multidominio, que le brinden al organismo húesped funciones distintas a aquellas que son debidas al dominio más conservado.

Estas limitantes fueron las que dieron cabida al desarrollo de una herramienta que permita analizar las proteínas a nivel dominio, considerando la similitud y su arquitectura y sin realizar asignaciones de funciones, dejando ésto último a criterio del investigador.

## 2.1. Hipótesis

Es posible estudiar la arquitectura de dominios en proteínas para mejorar su anotación, desarrollando una estrategia que pueda servir para cualquier proteína en cualquier organismo mediante el uso de herramientas de cómputo.

## 2.2. Objetivo general

Crear un software para la identificación de dominios funcionales en genomas y transcriptomas con el uso de Perl como lenguaje de programación, con la finalidad de estudiar la combinación y distribución de los dominios funcionales conocidos en proteínas.

## 2.3. Objetivos particulares

1. Analizar endonucleasas con la misma función en bacterias distintas para comparar sus dominios y estructuras. El software debe ser capaz de mostrar con las mismas características visuales aquellos dominios compartidos por todas las secuencias.

2. Analizar enzimas de restricción con funciones muy similares para comparar sus dominios y estructuras. El software debe diferenciar visualmente dominios distintos en todas las secuencias.

3. Analizar los dominios en proteínas con motivos Rossmann fold de unión a distintos dinucleótidos para observar su arquitectura a nivel de dominios. El software debe diferenciar de forma visual dominios similares.

4. Identificar dominios hidrolasas de distintas especies de *Bifidobacterium* en un metagenoma de pulque como base de datos local. El software debe ser capaz de identificar dominios en una serie de secuencias de entrada y posteriormente realizar una búsqueda de esas arquitecturas de dominios en un archivo local.

# Capítulo 3

# Métodos y materiales

## 3.1. Análisis de desarrollo

Después de usar las herramientas actuales de identificación de dominios funcionales y analizar sus resultados, se planeó la creación de un software que permita obtener resultados similares en un menor tiempo, con mayor disponibilidad, y sin etiquetar una proteína entera a una función específica.

Se requiere que el software reciba como entrada un archivo con una o más secuencias de proteínas con especificaciones FASTA, y que se produzca como salida un archivo con los dominios funcionales identificados en la(s) secuencia(s) (Figura 3.1).

Para realizar la identificación de los dominios, se hará uso de la *suite* de HMMER, que permite realizar perfiles haciendo uso de HMM.

### 3.1.1. Entradas

El sofware a desarrollar debe recibir dos o más de las siguentes entradas:

**Base de datos de dominios funcionales** Contiene modelos estadísticos (HMM)

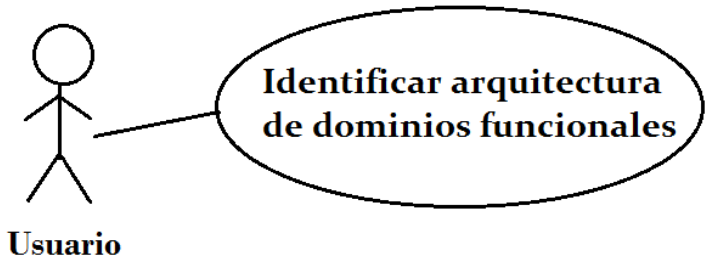



**Figura 3.1:** Diagrama de caso de uso.

con los nombres e identificadores de los dominios funcionales conocidos, que se usarán para etiquetar cada uno de los dominios encontrados.

**Archivo de texto con secuencias FASTA** Archivo con formato FASTA con las proteínas de las que se desea identificar los posibles dominios existentes.

**Archivo con anotaciones** Contiene datos de un genoma o transcriptoma con anotaciones sobre dominios funcionales. En este archivo se buscarán los dominios encontrados en el archivo FASTA.

**Archivo con arquitecturas de dominios** Contiene arquitecturas de dominios funcionales compuestas por su identificador y separados por comas. Cada línea en este archivo representa una arquitectura.

**Parámetros de ajuste** Valores que aumentan o reducen el espectro de resultados.

### 3.1.2. Procesos

El software deberá procesar todos los parámetros de entrada. El procesamiento de los datos es de la forma:

1. Establecer los parámetros de ajuste a los brindados por el usuario.
2. Leer todas las secuencias del archivo FASTA e identificar los dominios funcionales usando HMM.
3. Obtener los nombres e identificadores de los dominios encontrados realizando una búsqueda en la base de datos de dominios suministrada por el usuario.
4. Generar una tabla con los dominios encontrados para cada secuencia, brindando nombre, identificador, posición en la secuencia, *e-value* y *bitscore*, y descripción.
5. Generar una gráfica por cada secuencia en la que se muestren los dominios identificados, distiguiéndolos por colores.

### 3.1.3. Salidas

una vez que se haya ejecutado de forma exitosa, el software generará una o más de las siguientes entradas:

- Archivo `.table` con la descripción de los dominios encontrados.
- Archivo `.pdf` con las gráficas de todas las secuencias para las que se encontraron dominios.
- Archivo `.out` con las arquitecturas similares encontradas en el archivo con anotaciones.

### 3.1.4. Herramientas auxiliares al desarrollo

Este trabajo se realizó usando la base de datos Pfam (v29.0, actualmente está disponible la versión 30.0) para la identificación de dominios, esta base de datos se puede obtener del sitio de pfam `http://pfam.xfam.org/` en el apartado FTP->releases->Pfam29.0->Pfam-A.hmm.gz.

Para la creación y alineamiento de los perfiles HMM se hizo uso de `hmmscan` en la *suite* de HMMER. `hmmscan` genera de forma opcional, una tabla de resultados con distintas columnas de las que se seleccionaron solo aquellas columnas relevantes para este trabajo.

Las secuencias de aminoácidos de las proteínas analizadas fueron tomadas de la base de datos del NCBI en formato FASTA.

## 3.2. Pipeline

Se construyó el *pipeline* **SDA** (*Scan Domain Architecture*) para la identificación y búsqueda de arquitecturas de dominios. El *pipeline* fue desarrollado en Perl y cuenta con una interfaz gráfica desarrollada en Java que facilita su uso a quienes no están familiarizados con la terminal.

SDA puede ser usado de dos formas: primero, para estudiar la arquitectura de dominios de una o más secuencias de aminoácidos; segundo, para realizar búsquedas de dominios en una base de datos local en formato tabular con la anotación de los dominios en una de sus columnas. Antes de usar SDA, el usuario debe ejecutar `hmmpress <pathToPfam.hmm>` para generar cuatro archivos binarios que `hmmscan` usará para generar los perfiles.

Una vez que el usuario indica la ruta a la base de datos Pfam y los archivos con las secuencias a analizar, SDA trabaja en cuatro pasos (Figura 3.2):

1. HMMER crea el perfil de cada secuencia en un archivo FASTA ingresado usando Pfam-A como referencia.
2. Si se desean realizar búsquedas en un achivo local, SDA lee y procesa el archivo para para obtener todas las arquitecturas de dominios en ellas.
3. En caso de identificar dominios en el archivo FASTA con los parámetros indicados, se genera un archivo con una tabla resumen de HMMER y un archivo con las gráficas de los dominios encontrados para cada secuencia.
4. Si la búsqueda fue local y existen arquitecturas similares entre el archivo FASTA y el archivo local de anotaciones, se crea un tercer archivo con los resultados de las similitudes que incluye una tabla de frecuencias de cada dominio encontrado para observar su recurrencia.

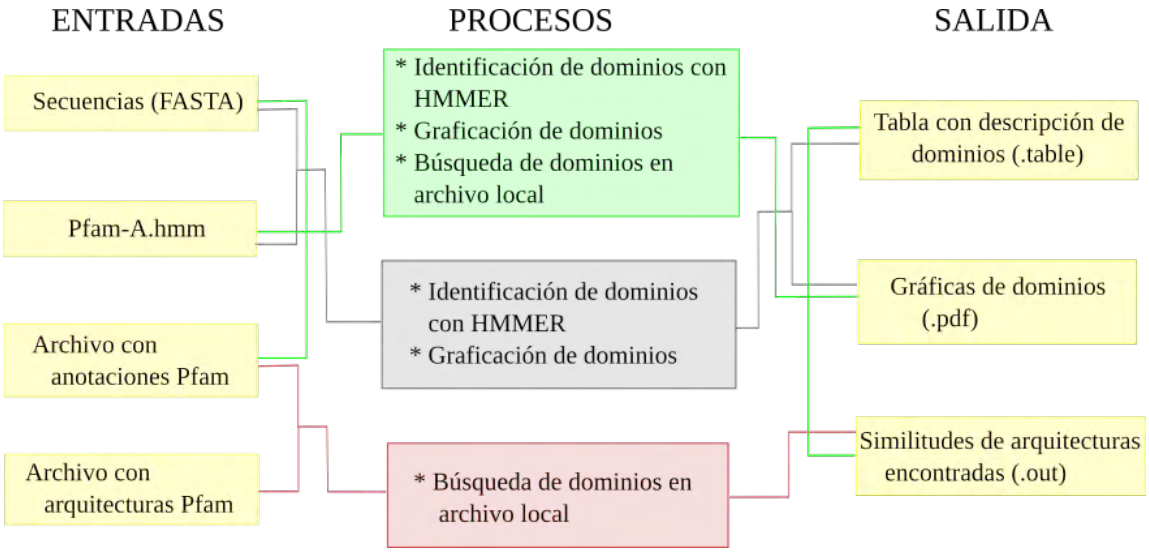



**Figura 3.2:** Entradas, procesos y salidas de SDA.

### 3.2.1. Preprocesamiento de archivos

Para que SDA funcione de forma correcta, los archivos de entrada deben cumplir ciertas características.

- El encabezado del archivo de anotaciones para la columna que contiene los dominios identificados debe comenzar con *pfam*. Si se trata de un genoma, el encabezado de la anotación del genoma debe contener *gene_id* y si es un transcriptoma *transcript*.

- El archivo de secuencias FASTA debe tener al menos una línea en blanco después de cada secuencia, incluso de la última secuencia pues `hmmscan` no funciona correctamente si ésto no se cumple.

Los identificadores de PFAM están conformados de la forma:

`PF`*`numero.version`*

donde `numero` es un número de 5 caracteres que identifica de forma única a ese dominio, y `version` es un número de 1 o dos caracteres, que se refiere a la versión del perfil. Debido a que con el tiempo se secuencian nuevas proteínas, los perfiles de dominios se actualizan con esas nuevas proteínas. Esos nuevos perfiles son una versión que representa una mayor cantidad de proteínas.

Todos los software que se dedican a analizar dominios funcionales, ignoran el número de versión del identificador PFAM. Del mismo modo, SDA ignora el número de versión del perfil y se enfoca sólo en el identificador.

### 3.2.2. Requerimientos

Para hacer uso de SDA se requiere lo siguiente:

- Sistema operativo Unix/Linux (SDA fue probado en sistemas basados en Ubuntu y basados en RHEL).
- Perl v5.18.12 o superior y módulo SVG (para las gráficas).
- La suite de HMMER (v3.1b1 o v3.1b2)[1] para la creación de los HMM.
- La base de datos de pfam Pfam-A.hmm (v29.0 o v30.0) que contiene los perfiles para la identificación de los dominios[2].
- Java 7 o superior (en caso de que se prefiera usar la interfaz gráfica).

### 3.2.3. Ajuste de parámetros

Al igual que cuando se realizan búsquedas de dominios en la base de datos Pfam, **SDA** permite ajustar parámetros como el *e-value* y la precisión con la finalidad de aumentar la permisividad de los resultados en cada ejecución. El valor *default* para el *e-value* es de $1e^{-}10$, mientras que para la precisión es de 0.85.

SDA también permite elegir la cantidad de procesadores a usar para la creación de los perfiles con HMMER.

### 3.2.4. Métrica de similitud

Cuando se buscan arquitecturas de dominios en un archivo local, se puntúan las similitudes de los dominios en la secuencia de entrada contra los dominios en el archivo local.

El puntaje de similitud $S_{SDA}$ entre las arquitecturas de dominios de dos secuencias se definió como

$$S_{SDA} = \frac{n}{\max\{l_1, l_2\}} \tag{3.1}$$

donde $n$ es la cantidad de dominios iguales compartidos por las dos secuencias, y $l_1, l_2$ son la cantidad de dominios encontrados en la secuencia uno y dos, respectivamente.

### 3.2.5. Resultados gráficos

En los resultados gráficos para los dominios identificados, se hace distinción con colores de hasta veinte dominios distintos por cada archivo de secuencias.

[1] `http://hmmer.org/`
[2] `ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/`

**Figura 3.3:** Representación de los dominios identificados en los resultados gráficos. Dominio completo a la izquierda, dominio incompleto a la derecha.

Los dominios se representan como rectángulos con bordes curvos si se identificaron de forma completa en la secuencia, y de rectánculos sin bordes curvos si no están completos (Figura 3.3).

SDA es de uso público y puede ser descargado desde `https://github.com/ramonflores/SDA`. El manual de usuario puede ser consultado en línea o en el apéndice de este trabajo.

## 3.3. Estructuras y alineamientos

Las estructuras 3D fueron generadas con RasMol 2.7.5.2 (`http://www.openrasmol.org/`)[66] con ficheros obtenidos del *Protein Data Bank* (PDB, `http://www.rcsb.org/pdb/home/home.do`)[67].

Para el caso de los alineamientos se usó Clustal Omega (`http://www.ebi.ac.uk/Tools/msa/clustalo/`), MUSCLE (`http://www.ebi.ac.uk/Tools/msa/muscle/`) y ClustalX.

# Capítulo 4

# Resultados

## 4.1. Caso I. REasas en bacterias

Para el primer caso de estudio se desea comparar REasas del tipo Mrr en bacterias con la finalidad de identificar los dominios asociados a esa enzima vital para la replicación de DNA dañado, además de presentar capacidad de aceptar DNA modificado de otro organismo. El Cuadro 4.1 lista las enzimas seleccionadas y el link al NCBI para obtener sus secuencias.

Se usó SDA para obtener y comparar los dominios en cada bacteria. Para este caso de estudio, SDA debe ser capaz de identificar con los mismos nombres, identificadores, y colores, aquellos dominios que sean idénticos en todas las secuencias. El Cuadro 4.2 muestra los datos obtenidos.

En la tabla de resultados se observa que todas las enzimas presentan dos dominios recurrentes: PF04471 identificado como *Mrr cat*, y PF14338, correspondiente a *Mrr N*, éste último, un dominio N-terminal; ésto se observa más claramente en los resultados gráficos (Figura 4.1).

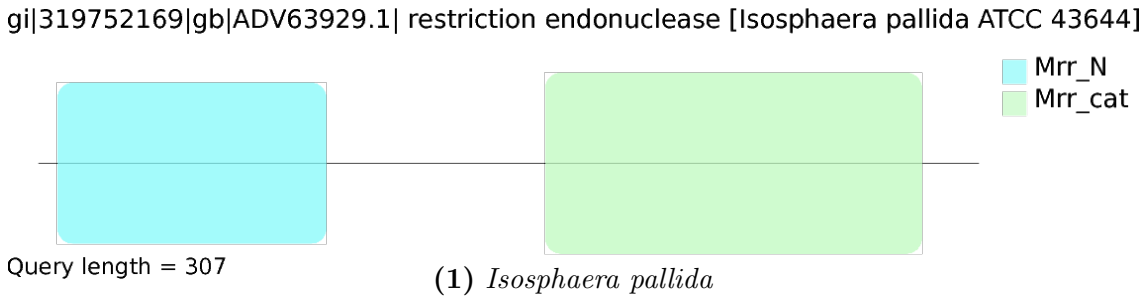
| # | **Organismo** | **Referencia** |
|---|---|---|
| 1 | *Isosphaera pallida* | `https://www.ncbi.nlm.nih.gov/protein/gi\|319752169` |
| 2 | *Sulfobacillus acidophilus* | `https://www.ncbi.nlm.nih.gov/protein/gi\|361053489` |
| 3 | *Cellulophaga lytica* | `https://www.ncbi.nlm.nih.gov/protein/gi\|324323260` |
| 4 | *Desulfobulbus propionicus* | `https://www.ncbi.nlm.nih.gov/protein/gi\|320123719` |
| 5 | *Turneriella parva* | `https://www.ncbi.nlm.nih.gov/protein/gi\|390612841` |
| 6 | *Desulfarculus baarsii* | `https://www.ncbi.nlm.nih.gov/protein/gi\|301640073` |
| 7 | *Crinalium epipsammum* | `https://www.ncbi.nlm.nih.gov/protein/gi\|428247692` |

**Cuadro 4.1:** Endonucleasas de restricción en distintas bacterias.

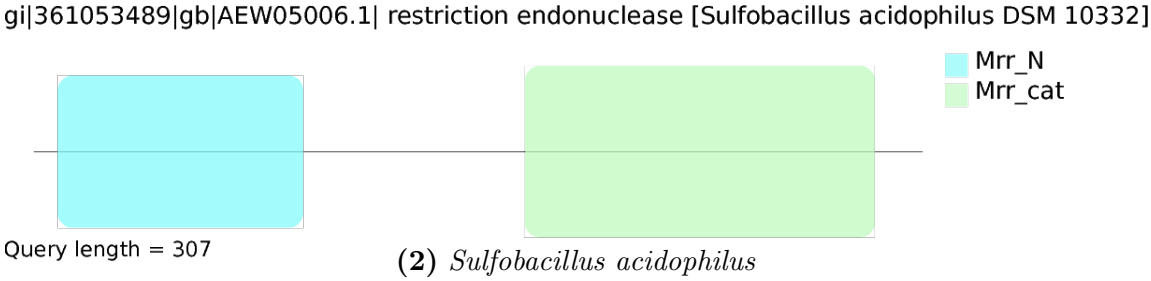
| query name | len | access. | target name | evalue | score | (ali) fr-to | (env) fr-to | acc | description of target |
|---|---|---|---|---|---|---|---|---|---|
| mrr.seq001 | 307 | PF04471 | Mrr cat | 2e-33 | 114.7 | 160-279 | 159-279 | 0.95 | Restriction endonuclease |
| mrr.seq001 | 307 | PF14338 | Mrr N | 5.4e-32 | 109.7 | 6 - 91 | 6 - 92 | 0.98 | Mrr N-terminal domain |
| mrr.seq002 | 307 | PF04471 | Mrr cat | 1.8e-32 | 111.7 | 164-281 | 160-281 | 0.96 | Restriction endonuclease |
| mrr.seq002 | 307 | PF14338 | Mrr N | 3e-21 | 75.3 | 8-90 | 7 - 91 | 0.96 | Mrr N-terminal domain |
| mrr.seq003 | 298 | PF04471 | Mrr cat | 3.4e-33 | 114.0 | 157-272 | 157-272 | 0.95 | Restriction endonuclease |
| mrr.seq003 | 298 | PF14338 | Mrr N | 1.9e-17 | 63.1 | 6 - 92 | 6 - 93 | 0.94 | Mrr N-terminal domain |
| mrr.seq004 | 303 | PF04471 | Mrr cat | 3.4e-36 | 123.6 | 159-277 | 157-277 | 0.94 | Restriction endonuclease |
| mrr.seq004 | 303 | PF14338 | Mrr N | 2.7e-31 | 107.4 | 6 - 91 | 6 - 92 | 0.99 | Mrr N-terminal domain |
| mrr.seq004 | 303 | PF13156 | Mrr cat 2 | 1.9e-12 | 47.0 | 191-282 | 182-289 | 0.90 | Restriction endonuclease |
| mrr.seq005 | 305 | PF04471 | Mrr cat | 9.3e-35 | 119.0 | 161-278 | 158-278 | 0.95 | Restriction endonuclease |
| mrr.seq005 | 305 | PF14338 | Mrr N | 3.1e-30 | 104.0 | 6 - 91 | 6 - 92 | 0.99 | Mrr N-terminal domain |
| mrr.seq005 | 305 | PF13156 | Mrr cat 2 | 1.5e-12 | 47.4 | 192-282 | 182-290 | 0.89 | Restriction endonuclease |
| mrr.seq006 | 308 | PF04471 | Mrr cat | 1.7e-34 | 118.2 | 163-282 | 162-282 | 0.94 | Restriction endonuclease |
| mrr.seq006 | 308 | PF14338 | Mrr N | 1.4e-29 | 102.0 | 6 - 90 | 6 - 92 | 0.98 | Mrr N-terminal domain |
| mrr.seq007 | 285 | PF04471 | Mrr cat | 1.3e-31 | 108.9 | 144-259 | 143-260 | 0.93 | Restriction endonuclease |
| mrr.seq007 | 285 | PF14338 | Mrr N | 1.4e-17 | 63.4 | 11 - 88 | 11 - 97 | 0.95 | Mrr N-terminal domain |

**Cuadro 4.2:** Resultados obtenidos para enzimas de restricción en bacterias. E-value = 1e-10, precisión = 0.85.

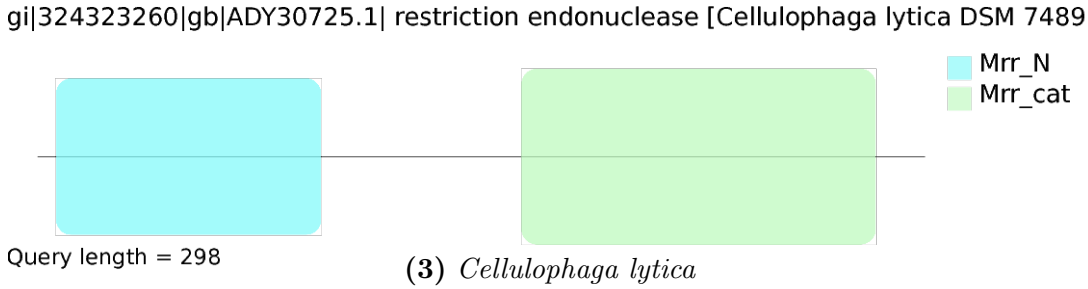
En un homólogo típico de *Mrr* se ha encontrado que hay al menos dos dominios: un dominio C-terminal altamente conservado semejante a la región catalítica de una endonucleasa, y un dominio N-terminal menos conservado que se cree es responsable del reconocimiento del DNA y activación del enlazamiento[68].



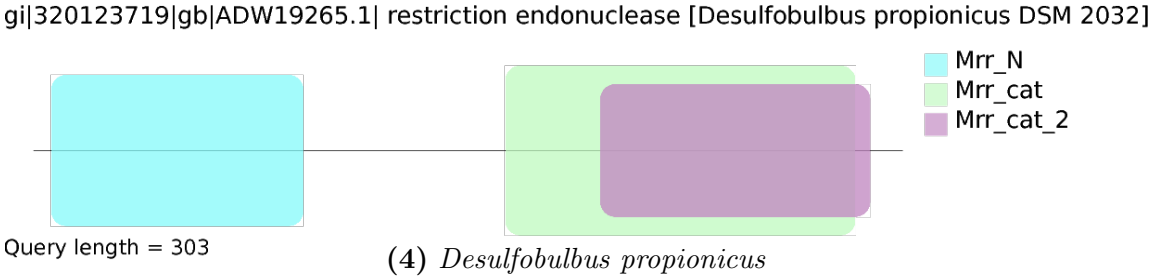

**(1)** *Isosphaera pallida*




**(2)** *Sulfobacillus acidophilus*

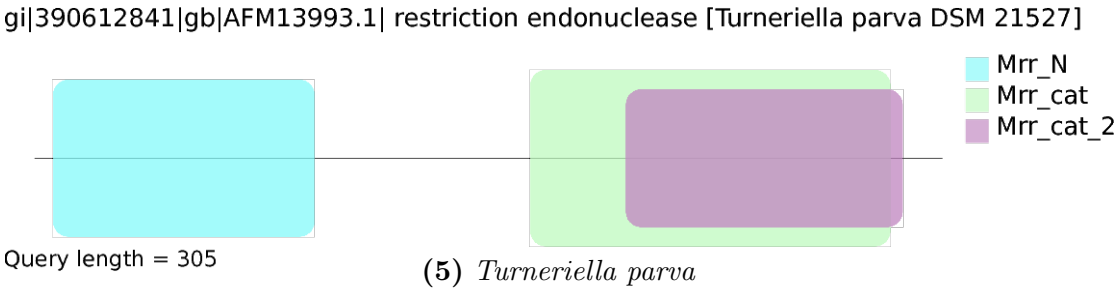



**(3)** *Cellulophaga lytica*

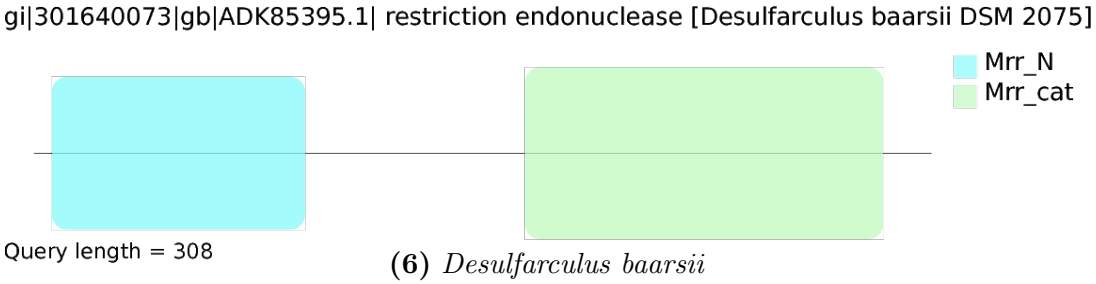



**(4)** *Desulfobulbus propionicus*

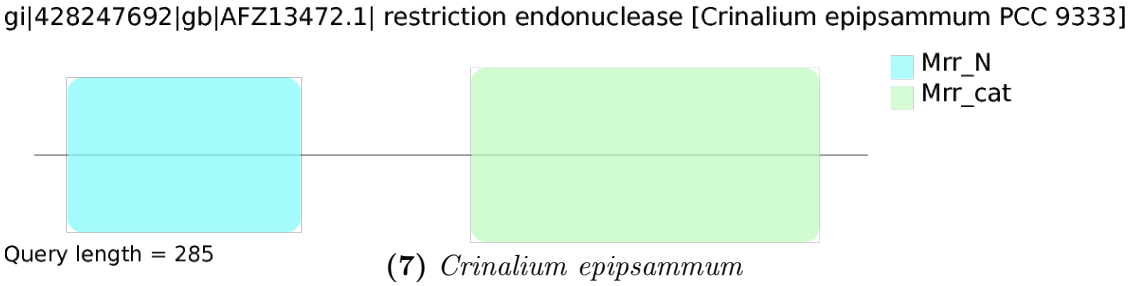
**Figura 4.1:** Dominios encontrados en endonucleasas de restricción en bacterias. e-value = $1e^{-10}$, precisión = 0.85, bitscore = 45




**(5)** *Turneriella parva*




**(6)** *Desulfarculus baarsii*




**(7)** *Crinalium epipsammum*

**Figura 4.1: (Continuación)** Dominios encontrados en endonucleasas de restricción en bacterias. e-value = $1e^{-10}$, precisión = 0.85, bitscore = 45

Los resultados obtenidos con SDA muestra que los *e-values* y *bit scores* de los dominios Mrr cat son mejores que los del dominio Mrr N, que puede interpretarse como que los primeros están más conservados que los segundos, coincidiendo así con lo reportado en la literatura. En las secuencias 4 y 5 se presenta el dominio adicional PF13156 identificado como *Mrr cat 2*, aunque está sobrepuesto con el dominio *Mrr cat*. A pesar de que su *e-value* y *score* son menos favorables, lo que debería hacer el usuario de SDA es realizar un análisis biológico para saber si la información obtenida tiene sentido biológico.

Para la estructura terciara se encontró que el identificador PDB para el dominio *Mrr cat* es 4F0P. La estructura para el dominio *Mrr N* es aún desconocida. La Figura 4.2 muestra la estructura del dominio *Mrr cat*.

SDA permite concluir que todas las enzimas tienen una estructura de dominios que está conservada en organismos distintos y algunos muy lejanos en un sentido evolutivo. La conservación se puede observar también en la estructura espacial.

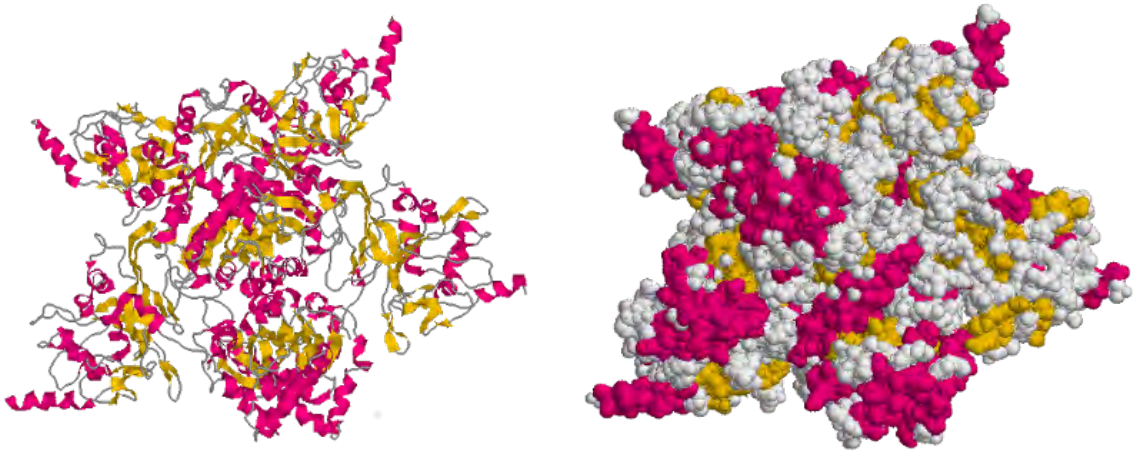


**Figura 4.2:** Estructura terciara del dominio PF04471 *Mrr cat* con identificador 4F0P en el PDB. Imagen generada con RasMol.

## 4.2. Caso II. REasas con sitios de reconocimiento similares

Anteriormente se mencionó que las enzimas son proteínas con capacidad catalítica cuya característica más sorprendente es su especificidad. Las enzimas de restricción reconocen una secuencia de nucleótidos y realizan esciciones dentro o fuera de esa secuencia. La base de datos REBASE (*Restriction Enzime Database*) brinda información como nombre, autor, tipo y secuencia de reconocimiento sobre enzimas[69].

De REBASE se obtuvieron los nombres de enzimas con sitios de reconocimiento idénticos o muy similares en con la finalidad de identificar y comparar los dominios y estructuras involucrados en esas enzimas. El sitio de reconocimiento elegido fue AAXXTT (donde X puede ser o G o C) y se desea observar si la variación de dos nucleotidos de DNA requiere de enzimas que tengan dominios funcionales diferentes. El Cuadro 4.3 lista las enzimas seleccionadas para el análisis.

| **Enzima** | **Rec/Esc** | **Referencia** |
|---|---|---|
| AclI | AA/CGTT | `http://rebase.neb.com/rebase/enz/AclI.html` |
| HindIII | A/AGCTT | `http://rebase.neb.com/rebase/enz/HindIII.html` |
| BpeI | AAGCTT | `http://rebase.neb.com/rebase/enz/BpeI.html` |
| Psp1406I | AA/CGTT | `http://rebase.neb.com/rebase/enz/Psp1406I.html` |
| M.LlaCl | AAGCTT | `http://rebase.neb.com/rebase/enz/M.LlaCI.html` |

**Cuadro 4.3:** Enzimas seleccionadas de REBASE. Rec/Esc muestra el sitio de reconocimiento, la diagonal indica que el sitio de escición es conocido.

Las secuencias de aminoácidos fueron obtenidas del NCBI. En el caso de la AclI, debido a que se encontraron proteínas ortólogas, se tomaron secuencias de tres organismos distintos. El Cuadro 4.4 lista las enzimas, organismos a que pertenecen y el enlace al NCBI.

| # | **Enzima** | **Organismo** | **Referencia** |
|---|---|---|---|
| 1 | AclI | *Staphylococcus aureus* | `https://www.ncbi.nlm.nih.gov/protein/604404848` |
| 2 | AclI | *Streptomyces galilaeus* | `https://www.ncbi.nlm.nih.gov/protein/16945721` |
| 3 | AclI | *Acinetobacter calcoaceticus* | `https://www.ncbi.nlm.nih.gov/protein/308229523` |
| 4 | HindIII | *Haemophilus influenzae* | `https://www.ncbi.nlm.nih.gov/protein/1174568` |
| 5 | BpeI | *Arabidopsis thaliana* | `https://www.ncbi.nlm.nih.gov/protein/gi\|332195477` |
| 6 | Psp1406I | *Helicobacter pylori* | `https://www.ncbi.nlm.nih.gov/protein/gi\|459509585` |
| 7 | M.LlaCL | *Lactococcus lactis* | `https://www.ncbi.nlm.nih.gov/protein/WP_014573488.1` |

**Cuadro 4.4:** Enzimas seleccionadas para el análisis con SDA.

Al archivo con las secuencias se aplicó SDA con los valores *default* para obtener los dominios asociados a las enzimas seleccionadas pero sólo se obtuvieron resultados para cuatro de las siete secuencias por lo que se volvió a ejecutar SDA con un *e-value* de 0,1, una precisión de 0,7 y un *bitscore* de 30. El Cuadro 4.5 muestra los resultados obtenidos y la Figura 4.3 las gráficas de los dominios.

SDA permite observar que la variación de dos nucleótidos en el DNA, requiere de tener enzimas de restricción con diferentes dominios para poder realizar la misma función. Esto sugiere que en la evolución se seleccionaron diferentes estrategias para poder realizar una función especializada.

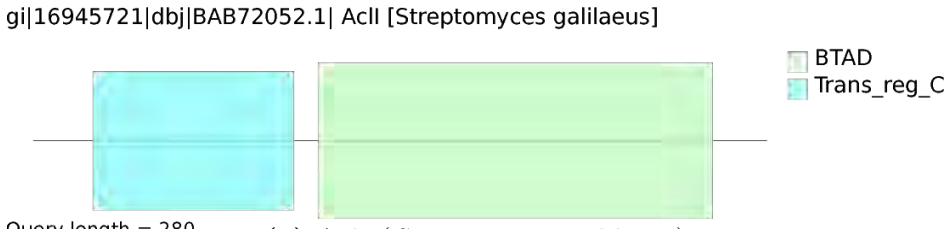
Lo siguiente es realizar la comparación de las estructuras para identificar si hay una relación entre la similitud de los sitios de reconocimiento y corte con la súperfamilia a la que una enzima pertenecen.

El Cuadro 4.6 muestra el dominio, su identificador en el PDB, el organismo al que pertenece la estructura terciaria y el sitio de donde se obtuvieron los archivos para la graficación. Las estructuras para cada dominio se muestran en la Figura 4.4.

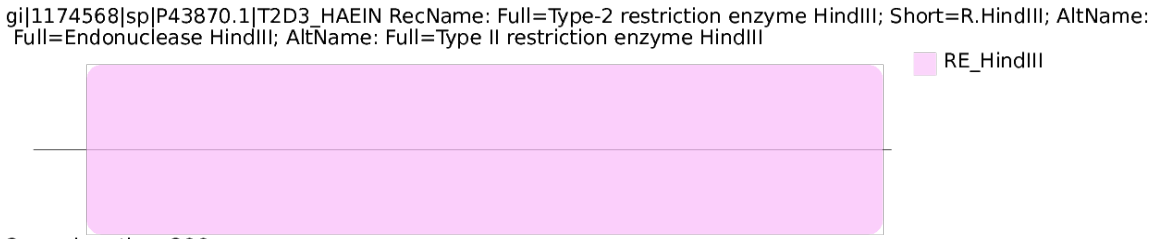
La diversidad funcional de la evolución en las proteínas analizadas, también se refleja en la estructura pues se puede observar que las estructuras también son distintas.

| query name | qlen | access. | target name | evalue | score | (ali) fr-to | (env) fr-to | acc | description of target |
|---|---|---|---|---|---|---|---|---|---|
| recSeq.seq002 | 280 | PF03704 | BTAD | 2.1e-42 | 144.9 | 105-250 | 105-250 | 0.99 | Bacterial transcriptional activator domain |
| | | PF00486 | Trans reg C | 5.8e-08 | 32.6 | 22 - 96 | 17 - 98 | 0.80 | Transcriptional regulatory protein, C terminal |
| recSeq.seq004 | 300 | PF09518 | RE HindIII | 3.1e-53 | 180.7 | 18 - 287 | 4 - 293 | 0.92 | HindIII restriction endonuclease |
| recSeq.seq005 | 343 | PF00010 | HLH | 3.1e-08 | 33.3 | 146-193 | 143 - 193 | 0.95 | Helix-loop-helix DNA-binding domain |
| recSeq.seq006 | 125 | PF01042 | Ribonuc L-PSP | 1.7e-42 | 144.1 | 10 - 124 | 7 - 125 | 0.98 | Endoribonuclease L-PSP |
| recSeq.seq007 | 296 | PF01555 | N6 N4 Mtase | 1.4e-46 | 158.9 | 23 - 287 | 23 - 290 | 0.89 | DNA methylase |

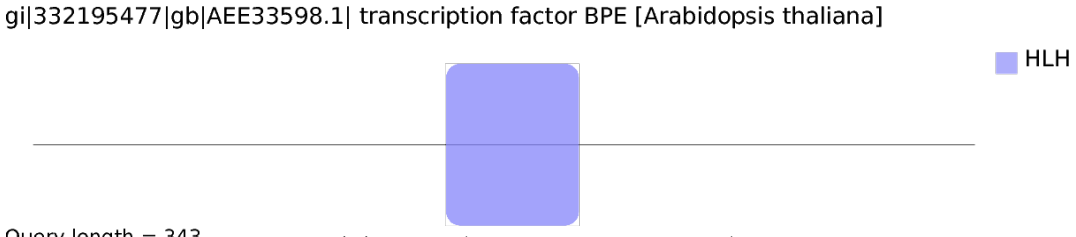
**Cuadro 4.5:** Tabla de resultados de SDA en enzimas con sitio de reconocimiento similar.



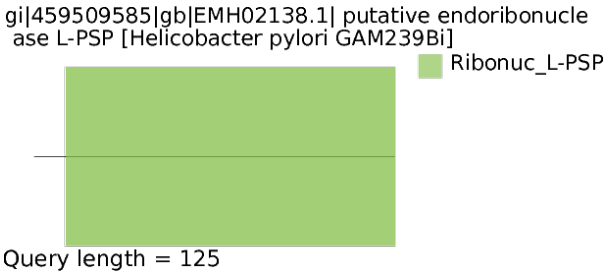

**(2)** AclI (*Streptomyces galilaeus*)



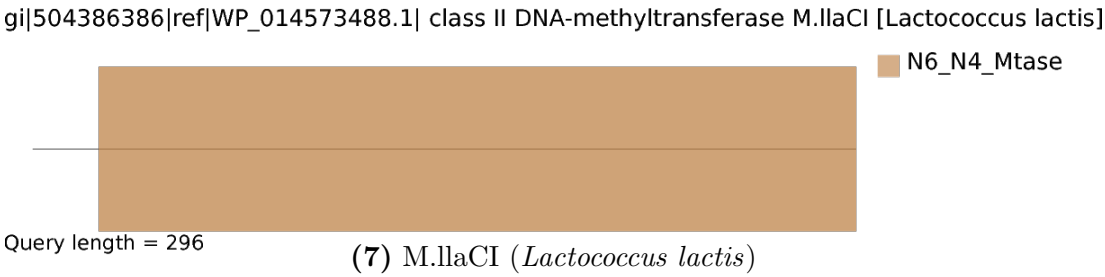

**(4)** HindIII (*Haemophilus influenzae*)




**(5)** BpeI (*Arabidopsis thaliana*)




**(6)** Psp1406I (*Helicobacter pylori*)



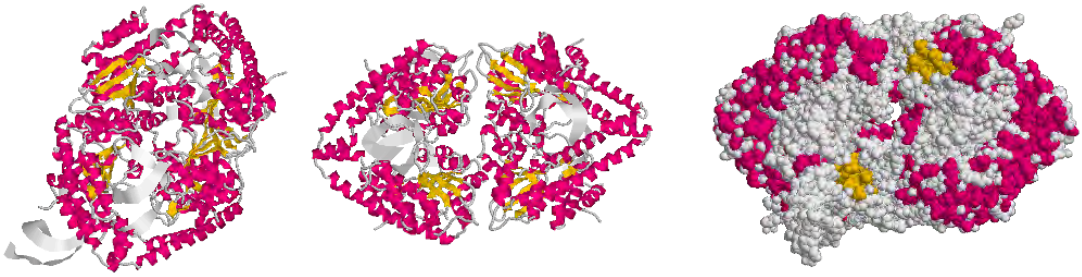

**(7)** M.llaCI (*Lactococcus lactis*)

**Figura 4.3:** Dominios encontrados en enzimas con sitios de reconocimiento similares.

| # | Dominio | PDB | Organismo | Referencia |
|---|---|---|---|---|
| 2 | BTAD / Trans reg C | 2FF4 | *Mycobacterium tuberculosis* | `http://www.rcsb.org/pdb/explore/explore.do?structureId=2FF4` |
| 4 | RE HindIII | 2E52 | *Haemophilus influenzae* | `http://www.rcsb.org/pdb/explore/explore.do?structureId=2E52` |
| 5 | HLH | 2QL2 | *Mus musculus* | `http://www.rcsb.org/pdb/explore/explore.do?structureId=2ql2` |
| 6 | Ribonuc L-PSP | 1QU9 | *Escherichia coli* | `http://www.rcsb.org/pdb/explore/explore.do?structureId=1QU9` |
| 7 | N6 N4 Mtase | 1G60 | *Moraxella bovis* | `http://www.rcsb.org/pdb/explore/explore.do?structureId=1G60` |

**Cuadro 4.6:** Dominio, identificador y fuente para las REasas seleccionadas.

**(2)** 2FF4 - BTAD



**(4)** 2E52 - RE HindIII

**(5)** 2QL2 - HLH

**Figura 4.4:** Estructura terciaria de los dominios encontrados con SDA en REasas con sitios de reconocimiento similares. Estructuras generadas con RasMol.

**(6)** 1QU9 - Ribonuc L-PSP

**(7)** 1G60 - N6 N4 Mtase

**Figura 4.4: (Continuación).** Estructura terciaria de los dominios encontrados con SDA en REasas con sitios de reconocimiento similares. Estructuras generadas con RasMol.

## 4.3. Caso III. Enzimas con motivos Rossmann fold

Del artículo de Bottoms *et al.*[70] se tomaron algunas enzimas con motivos Rossmann fold para identificar los dominios asociados a estas enzimas en búsqueda de similitudes estructurales. El Cuadro 4.7 lista las enzimas seleccionadas, el organismo al que pertenece y la fuente de donde se obtuvieron las secuencias.

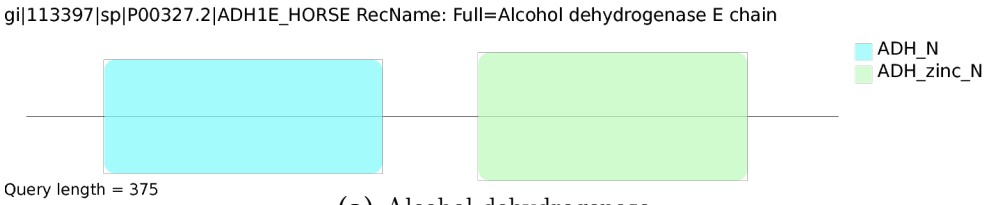
| # | **Enzima** | **Organismo** | **Referencia** |
|---|---|---|---|
| 1 | Alcohol dehydrogenase | *Equus caballus* | www.ncbi.nlm.nih.gov/protein/P00327.2 |
| 2 | GAPDH | *Escherichia coli* | www.ncbi.nlm.nih.gov/protein/ACI83891.1 |
| 3 | D-2-hydroxyisocaproate dehydrogenase | *Corynebacterium glutamicum* | www.ncbi.nlm.nih.gov/protein/BAA13523.1 |
| 4 | NADPH-flavin reductase | *Homo sapiens* | www.ncbi.nlm.nih.gov/protein/BAA05370.1 |
| 5 | Carbonyl reductase | *Mus musculus* | www.ncbi.nlm.nih.gov/protein/BAA05120.1 |
| 6 | Glucose oxidase | *Penicillium amagasakiense* | www.ncbi.nlm.nih.gov/protein/AAD01493.1 |
| 7 | Sarcosine oxidase | *Bacillus sp.* | www.ncbi.nlm.nih.gov/protein/BAA01410.1 |
| 8 | Carbohydrate kinase | *Escherichia coli* | www.ncbi.nlm.nih.gov/protein/WP_001298687.1 |

**Cuadro 4.7:** Enzimas con dominios Rossmann fold.

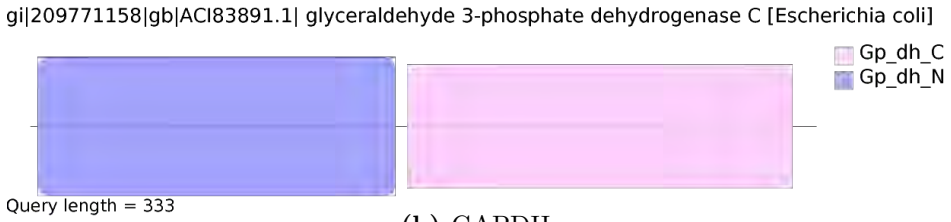
Los resultados de SDA se muestran en el Cuadro 4.8 y en la Figura 4.5.

| query name | len | access. | target name | evalue | score | (ali) fr-to | (env) fr-to | acc | description of target |
|---|---|---|---|---|---|---|---|---|---|
| rossmann. seq001 | 375 | PF00107 | ADH zinc N | 3.1e-25 | 88.4 | 203-324 | 203-337 | 0.93 | Zinc-binding dehydroge-nase |
| rossmann. seq001 | 375 | PF08240 | ADH N | 2.5e-24 | 85.1 | 35-160 | 34-161 | 0.90 | Alcohol dehydrogenase GroES-like domain |
| rossmann. seq002 | 333 | PF02800 | Gp dh C | 7.5e-61 | 204.2 | 155-313 | 155-313 | 0.98 | Glyceraldehyde 3-phosphate dehydr, C-terminal domain |
| rossmann. seq002 | 333 | PF00044 | Gp dh N | 6.1e-53 | 178.8 | 3-150 | 2-150 | 0.98 | Glyceraldehyde 3-phosphate dehydr, NAD binding domain |
| rossmann. seq003 | 320 | PF16654 | DAPDH C | 2.5e-67 | 225.4 | 118-268 | 118-269 | 0.99 | Diaminopimelic acid dehydrogenase C-terminal domain |
| rossmann. seq004 | 206 | PF13460 | NAD binding 10 | 8.5e-38 | 130.0 | 10-191 | 10-192 | 0.90 | NAD(P)H-binding |
| rossmann. seq005 | 244 | PF13561 | adh short C2 | 2.1e-55 | 187.7 | 14-241 | 14-242 | 0.96 | Enoyl-(Acyl carrier pro-tein) reductase |
| rossmann. seq005 | 244 | PF00106 | adh short | 1e-53 | 181.6 | 10-193 | 9-195 | 0.98 | short chain dehydrogena-se |
| rossmann. seq006 | 605 | PF00732 | GMC oxred N | 3.5e-73 | 246.4 | 43-352 | 43-353 | 0.95 | GMC oxidoreductase |
| rossmann. seq006 | 605 | PF05199 | GMC oxred C | 3.2e-23 | 82.6 | 452-592 | 452-593 | 0.85 | GMC oxidoreductase |
| rossmann. seq007 | 390 | PF01266 | DAO | 5.9e-52 | 177.2 | 6-361 | 6-361 | 0.85 | FAD dependent oxidore-ductase |
| rossmann. seq008 | 510 | PF01256 | Carb kinase | 4.9e-83 | 278.2 | 253-491 | 252-491 | 0.99 | Carbohydrate kinase |
| rossmann. seq008 | 510 | PF03853 | YjeF N | 3.6e-44 | 150.5 | 35-194 | 33-194 | 0.98 | YjeF-related protein N-terminus |

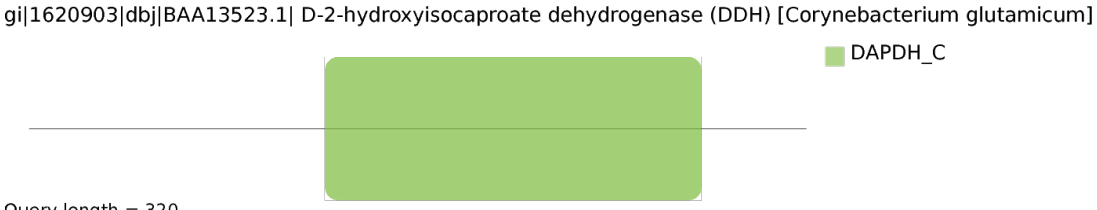
**Cuadro 4.8:** Resultados de SDA para enzimas con dominios Rossmann fold.



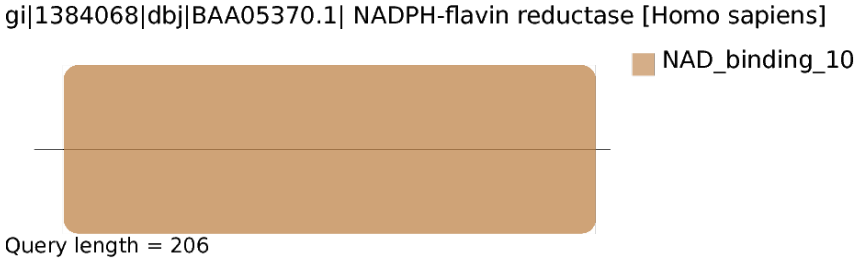

**(a)** Alcohol dehydrogenase



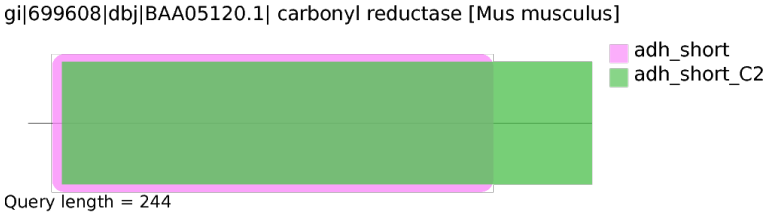

**(b)** GAPDH



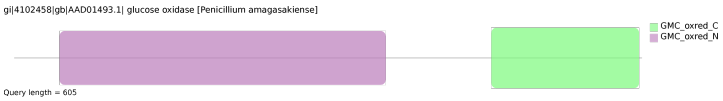

**(c)** D-2-hydroxyisocaproate dehydrogenase




**(d)** NADPH-flavin reductase

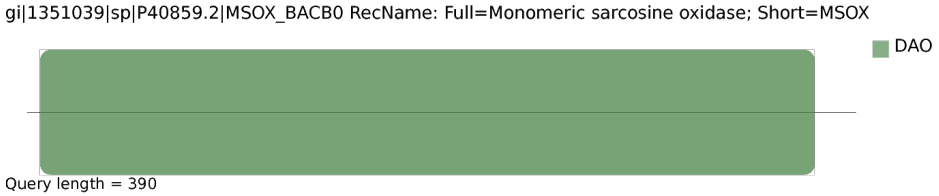



**(e)** Carbonyl reductase

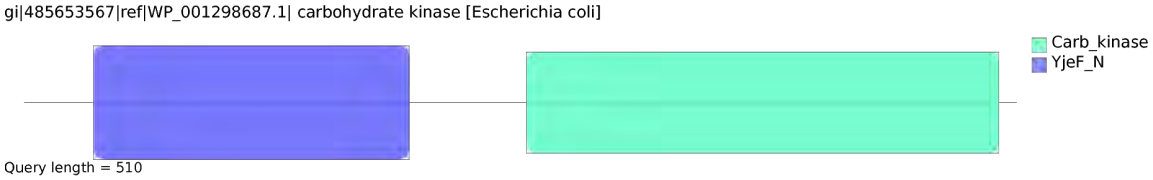



**(f)** Glucose oxidase

**Figura 4.5:** Resultados gráficos de SDA en proteínas con dominios Rossmann fold.




**(g)** Sarcosine oxidase




**(h)** Carbohydrate kinase

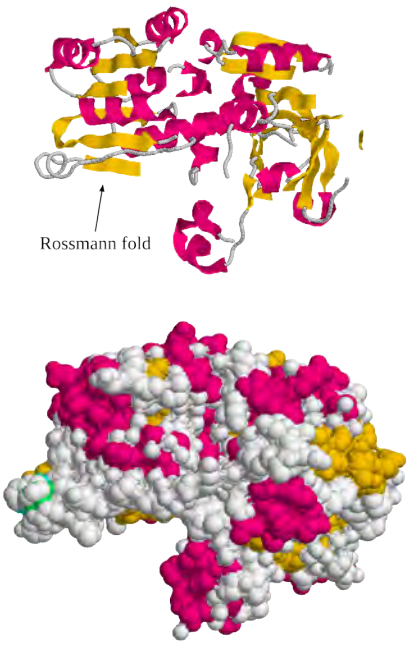
**Figura 4.5:** Resultados gráficos de SDA en proteínas con dominios Rossmann fold (continuación).

Los archivos de las estructuras, sus identificadores PDB y sus enlaces se muestran en el Cuadro 4.9. La Figura 4.6 muestra las estructuras graficadas.

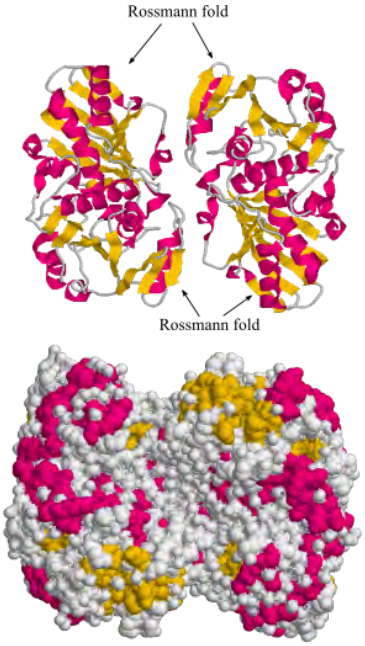
| # | Dominio | PDB | Organismo | Referencia |
|---|---|---|---|---|
| 1 | ADH zinc N / ADH N | 4OH1 | *[Clostridium] scindens* | `http://www.rcsb.org/pdb/explore/explore.do?structureId=4OH1` |
| 2 | Gp dh C / Gp dh N | 3PYM | *Saccharomyces cerevisiae* | `http://www.rcsb.org/pdb/explore/explore.do?structureId=3PYM` |
| 3 | DAPDH C | 1DAP | *Corynebacterium glutamicum* | `http://www.rcsb.org/pdb/explore/explore.do?structureId=1dap` |
| 4 | NAD biding 10 | 4R01 | *Streptococcus pneumoniae* | `http://www.rcsb.org/pdb/explore/explore.do?structureId=4r01` |
| 5 | adh short C2 | 1PR9 | *Homo sapiens* | `http://www.rcsb.org/pdb/explore/explore.do?structureId=1pr9` |
| 6 | GMC oxred N / GMC oxred C | 1GPE | *Penicillium amagasakiense* | `http://www.rcsb.org/pdb/explore/explore.do?structureId=1gpe` |
| 7 | DAO | 1EL5 | *Bacillus sp. B-0618* | `http://www.rcsb.org/pdb/explore/explore.do?structureId=1EL5` |
| 8 | Carb kinase / YjeF N | 3K5W | *Helicobacter pylori* | `http://www.rcsb.org/pdb/explore/explore.do?structureId=3k5w` |

**Cuadro 4.9:** Dominios, identificador y fuente para las secuencias analizadas.

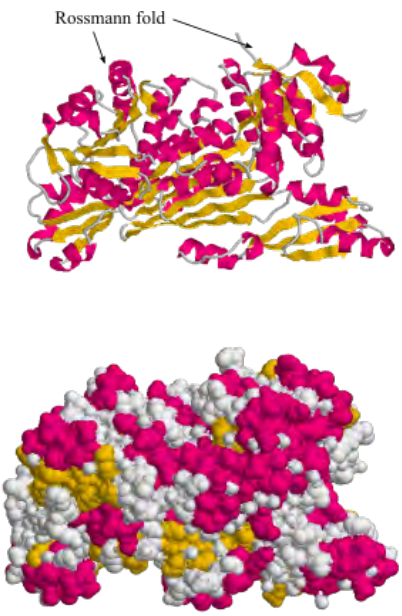
Para este caso de estudio, se observa que las diferentes enzimas presentan diferentes dominios funcionales. Sin embargo, esto no altera la estructura, la cual se conserva entre las distintas enzimas.



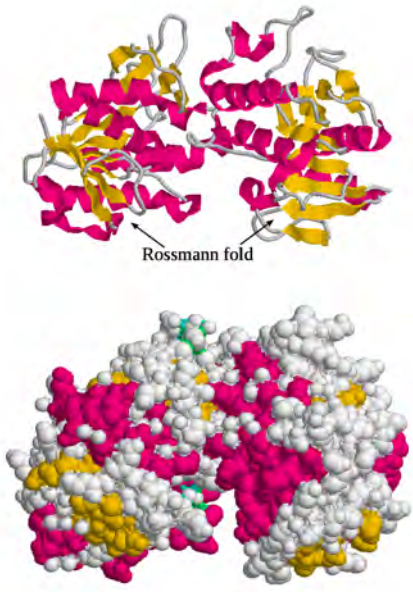

**(1)** 4OH1 para ADH zinc N / ADH N




**(2)** 3PYM para Gp dh C / Gp dh N

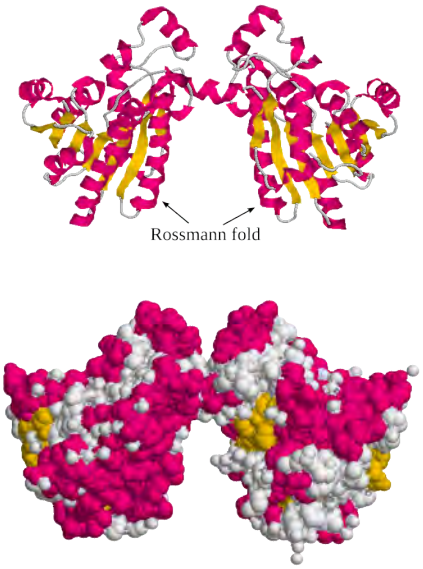



**(3)** 1DAP para DAPDH C

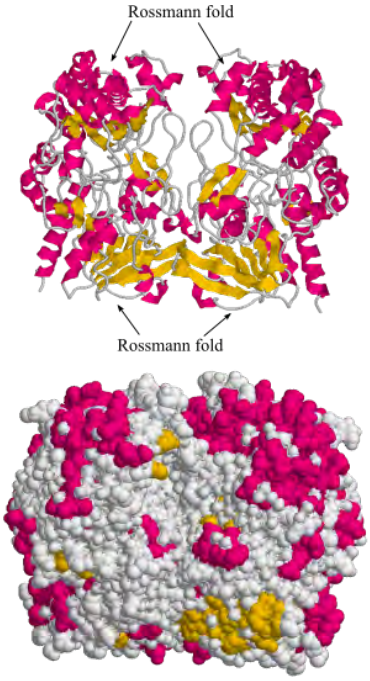



**(4)** 4R01 para NAD binding 10

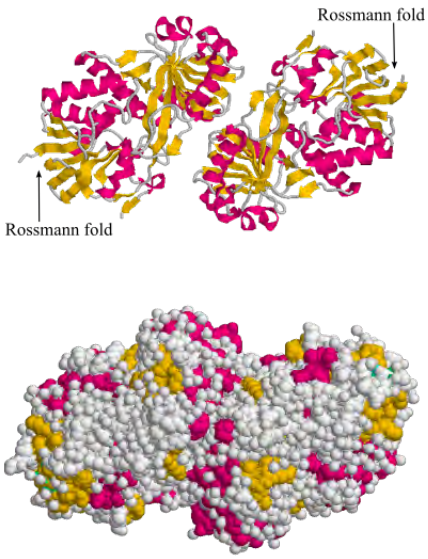
**Figura 4.6:** Estructuras de las proteínas con motivos Rossmann fold graficadas con RasMol.



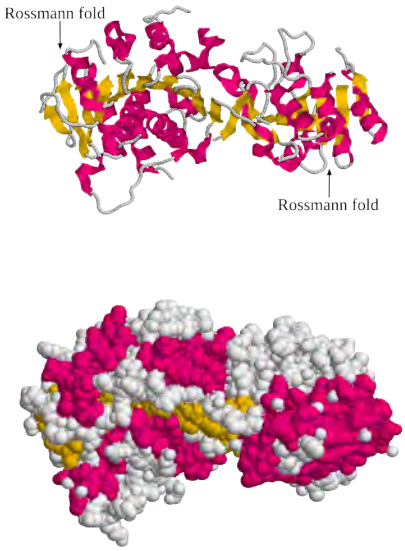

**(5)** 1PR9 para adh short C2




**(6)** 1GPE para GMC oxred N / GMC oxred C




**(7)** 1EL5 para DAO




**(8)** 3K5W para Carb kinase / YjeF N

**Figura 4.6: Continuación.** Estructuras de las proteínas con motivos Rossmann fold graficadas con RasMol.

## 4.4. Caso IV. Metagenoma del pulque

Si un usuario desea buscar arquitecturas de dominios en archivos locales, no le será posible hacer uso de herramientas como CDD, SMART o Pfam, por lo que deberá hacerlo directamente en el archivo o creando un programa que lo haga por él. Ésto, por supuesto, es una desventaja para quienes no cuentan con conocimientos de programación. SDA tiene la capacidad de permitir al usuario elegir un archivo "local" en formato tabular para realizar las búsquedas de arquitecturas.

Para este caso de estudio, se cuenta con el metagenoma del pulque en formato tabular y se desea identificar los dominios asociados a glicosil hidrolasas (EC 3.2.1.x), encargadas de descomponer y digerir celulosa y almidón[71][72], catalizar azúcares complejos para convertirlos en compuestos más sencillos que puedan ser asimilados por los humanos[73], y desempeñar funciones en procesos biológicos de vital importancia[74][75].

Se desea que los dominios de interés estén altamente relacionados a tres especies de *Bifidobacterium*. Las bifidobacterias bacterias son bacterias presentes en el tracto intestinal de algunos mamíferos que aunque no están en cantidades abundantes, son indispensables para la salud del tracto digestivo[76][77]. Para este estudio se consideran las especies *B. animalis*, *B. breve*, y *B. longum*. En distintos estudios se ha encontrado que mejoran la respuesta inmune a ciertas infecciones[78], reducen los tiempos de tránsito fisiológicos en el tracto intestinal[79][80], y mejoran la flora intestinal para salud humana[81][82].

El Cuadro 4.10 muestra las enzimas de interés, el organismo al que corresponden, y la referencia para obtener la secuencia.

| # | **Enzima** | **Organismo** | **Referencia** |
|---|---|---|---|
| 1 | Sucrose-6-phosphate hydrolase | *B. animalis* | `https://www.ncbi.nlm.nih.gov/protein/549638663` |
| 2 | Glycosyl hydrolase | *B. longum* | `https://www.ncbi.nlm.nih.gov/protein/WP_007052488.1` |
| 3 | Sucrose-6-phosphate hydrolase | *B. breve* | `https://www.ncbi.nlm.nih.gov/protein/WP_019727981.1` |

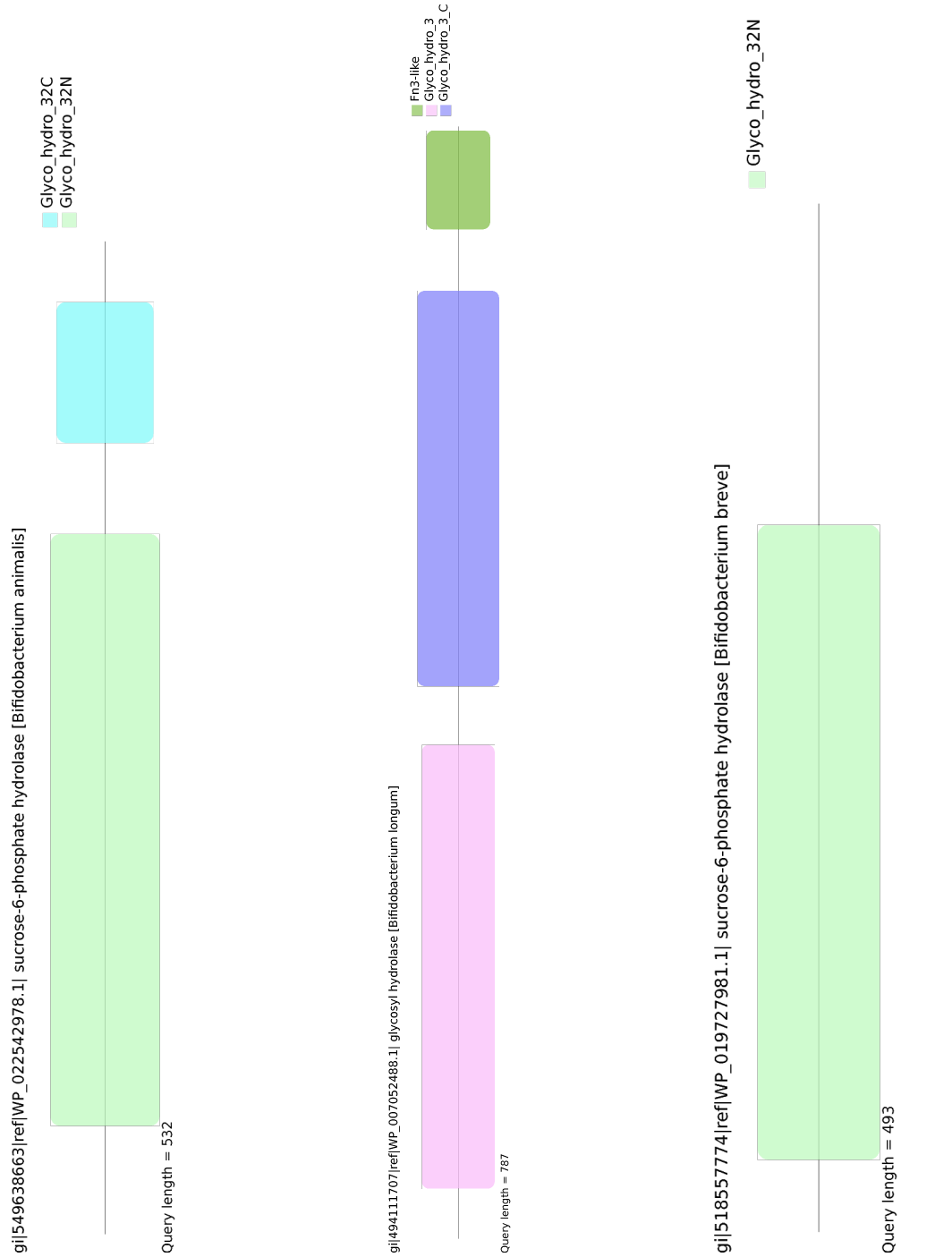
**Cuadro 4.10:** Enzimas usadas para el análisis de SDA para dominios Rossmann fold.

Los resultados de SDA para las glicosil hidrolasas se muestran en el Cuadro 4.11 y los resultados gráficos en la Figura 4.7.

Así mismo, el Cuadro 4.12 lista los dominios funcionales encontrados, su identificador en el PDB, y el enlace al fichero para obtener las gráficas de sus estructuras superiores (Figura 4.8).

| query name | len | access. | target name | evalue | score | (ali) fr-to | (env) fr-to | acc | description of target |
|---|---|---|---|---|---|---|---|---|---|
| pulque.seq001 | 532 | PF00251 | Glyco hydro 32N | 1.3e-90 | 303.9 | 57-368 | 57-368 | 0.93 | Glycosyl hydrolases family 32 N-terminal domain |
| pulque.seq001 | 532 | PF08244 | Glyco hydro 32C | 1.1e-13 | 51.6 | 416-490 | 416-490 | 0.88 | Glycosyl hydrolases family 32 C terminal |
| pulque.seq002 | 787 | PF00933 | Glyco hydro 3 | 3.2e-69 | 233.6 | 35-345 | 35-346 | 0.95 | Glycosyl hydrolase family 3 N terminal domain |
| pulque.seq002 | 787 | PF01915 | Glyco hydro 3 C | 5.5e-42 | 143.8 | 386-662 | 385-662 | 0.91 | Glycosyl hydrolase family 3 C-terminal domain |
| pulque.seq002 | 787 | PF14310 | Fn3-like | 6e-18 | 64.6 | 705-774 | 705-774 | 0.97 | Fibronectin type III-like domain |
| pulque.seq003 | 493 | PF00251 | Glyco hydro 32N | 6.6e-97 | 324.6 | 34-332 | 34-335 | 0.93 | Glycosyl hydrolases family 32 N-terminal domain |

**Cuadro 4.11:** Resultados de SDA para las enzimas del pulque.




**(a)** Sucrose-6-phosphate hydrolase (*B. animalis*)

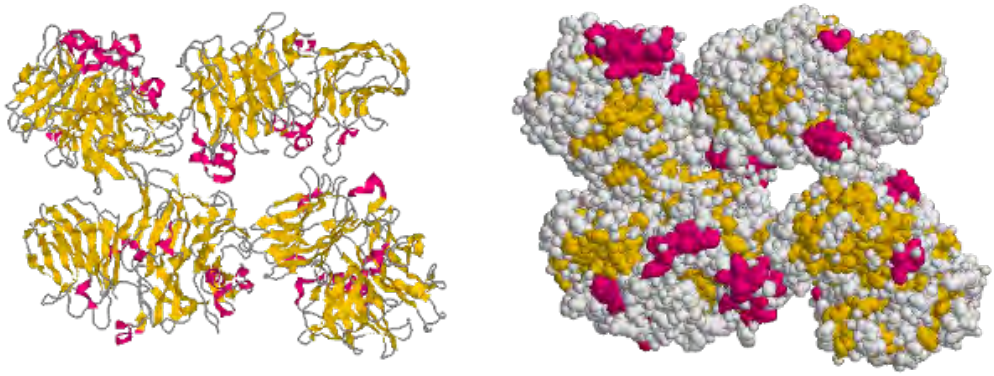
**(b)** Glycosyl hydrolase (*B. longum*)

**(c)** Sucrose-6-phosphate hydrolase (*B. breve*)

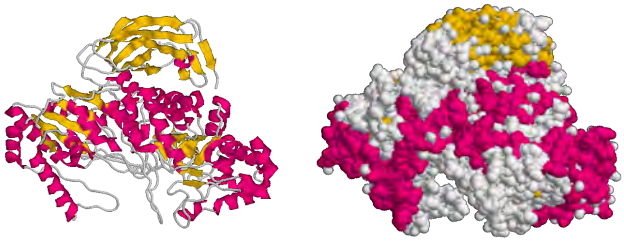
**Figura 4.7:** Dominios de hidrolasas encontrados en las secuencias de los *Bifidobacterium*.

| # | Dominio | PDB | Organismo | Referencia |
|---|---|---|---|---|
| 1 | Glyco hydro 32N / Glyco hydro 32C | 4FFF | *Paenarthrobacter ureafaciens* | `http://www.rcsb.org/pdb/explore/explore.do?structureId=4fff` |
| 2 | Glyco hydro 3 / Glyco hydro 3 C / Fn3-like | 2X41 | *Thermotoga neapolitana* | `http://www.rcsb.org/pdb/explore/explore.do?structureId=2x41` |
| 3 | Glyco hydro 32N | 3RWK | *Aspergillus ficuum* | `http://www.rcsb.org/pdb/explore/explore.do?structureId=3rwk` |

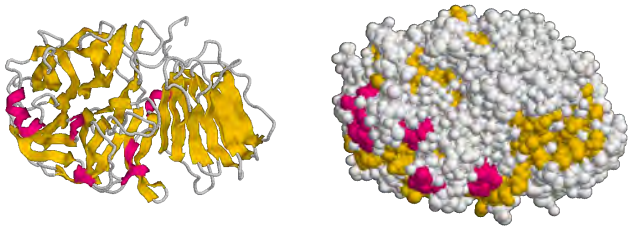
**Cuadro 4.12:** Dominios, identificador y fuente para las secuencias de *Bifidobacterium*.



**(1)** 4FFF para Glyco hydro 32N / Glyco hydro 32C



**(2)** 2X41 para Glyco hydro 3 / Glyco hydro 3 C / Fn3-like



**(3)** 3RWK para Glyco hydro 32N

**Figura 4.8:** Estructuras de las enzimas de *Bifidobacterium* con dominios hidrolasas.

Para este caso de estudio se contó también con una lista de anotación para el metagenoma del pulque, por lo que SDA arroja los resultados de los dominios en las enzimas buscadas y sus similitudes con los dominios en la lista de anotación. Los resultados se muestran a continuación:

```
# Pfam INPUT sequence for gi|549638663|ref|WP_022542978.1| sucrose-6-phosphate hydrolase
  [Bifidobacterium animalis]
# -- ALI --  -- ENV --
# BEGIN END  BEGIN END  PFAM_ID     SHORT_NAME        E-VALUE  DESCRIPTION OF TARGET
    57  368    57  368  PF00251     Glyco_hydro_32N   9.6e-91  Glycosyl hydrolases family
                                                               32 N-terminal domain
   416  490   416  490  PF08244     Glyco_hydro_32C   1.1e-13  Glycosyl hydrolases family
                                                               32 C terminal
## Pfam structure SIMILARITY for GENE: comp1218_c0
    14  326    --   --  PF00251.15  Glyco_hydro_32N   4.4e-88  Glycosyl hydrolases family
                                                               32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp13476_c0
    32  125    --   --  PF00251.15  Glyco_hydro_32N   1.1e-39  Glycosyl hydrolases family
                                                               32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp13900_c0
    76  277    --   --  PF00251.15  Glyco_hydro_32N   8.7e-27  Glycosyl hydrolases family
                                                               32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp13901_c0
     4   82    --   --  PF00251.15  Glyco_hydro_32N   4.4e-07  Glycosyl hydrolases family
                                                               32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp16517_c0
     5  171    --   --  PF00251.15  Glyco_hydro_32N   3.7e-19  Glycosyl hydrolases family
                                                               32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp17480_c0
    76  217    --   --  PF00251.15  Glyco_hydro_32N   3.4e-21  Glycosyl hydrolases family
                                                               32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp23569_c0
     7  153    --   --  PF00251.15  Glyco_hydro_32N   2.4e-21  Glycosyl hydrolases family
                                                               32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp23587_c0
     7  161    --   --  PF00251.15  Glyco_hydro_32N   4.4e-21  Glycosyl hydrolases family
                                                               32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp27727_c0
    10   96    --   --  PF00251.15  Glyco_hydro_32N     3e-05  Glycosyl hydrolases family
                                                               32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp28146_c0
    10   94    --   --  PF00251.15  Glyco_hydro_32N   7.8e-19  Glycosyl hydrolases family
                                                               32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp30085_c0
     5  117    --   --  PF00251.15  Glyco_hydro_32N   3.1e-11  Glycosyl hydrolases family
                                                               32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp6771_c0
    10  322    --   --  PF00251.15  Glyco_hydro_32N   1.5e-23  Glycosyl hydrolases family
                                                               32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp8872_c0
    21  334    --   --  PF00251.15  Glyco_hydro_32N   2.3e-23  Glycosyl hydrolases family
                                                               32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp30828_c0
     7  106    --   --  PF00251.15  Glyco_hydro_32N   2.2e-28  Glycosyl hydrolases family
                                                               32 N-terminal domain
    39  112    --   --  PF04616.9   Glyco_hydro_43    1.2e-05  Glycosyl hydrolases family 43
## Pfam structure SIMILARITY for GENE: comp2468_c0
    45  349    --   --  PF00251.15  Glyco_hydro_32N   8.9e-82  Glycosyl hydrolases family
                                                               32 N-terminal domain
    63  240    --   --  PF04616.9   Glyco_hydro_43    2.4e-05  Glycosyl hydrolases family 43
   388  460    --   --  PF08244.7   Glyco_hydro_32C   1.1e-05  Glycosyl hydrolases family
                                                               32 C terminal
```

```
## Pfam structure SIMILARITY for GENE: comp14192_c0
    33  343    --   --  PF00251.15  Glyco_hydro_32N   3.5e-97  Glycosyl hydrolases family
                                                                 32 N-terminal domain
   383  458    --   --  PF08244.7   Glyco_hydro_32C     1e-13  Glycosyl hydrolases family
                                                                 32 C terminal
## Pfam structure SIMILARITY for GENE: comp14228_c0
    33  343    --   --  PF00251.15  Glyco_hydro_32N  1.4e-100  Glycosyl hydrolases family
                                                                 32 N-terminal domain
   382  458    --   --  PF08244.7   Glyco_hydro_32    5.7e-14  Glycosyl hydrolases family
                                                                 32 C terminal
## Pfam structure SIMILARITY for GENE: comp13535_c0
   186  624    --   --  PF02435.11  Glyco_hydro_68   9.8e-106  Levansucrase/Invertase
   311  428    --   --  PF00251.15  Glyco_hydro_32N   0.00019  Glycosyl hydrolases family
                                                                 32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp15259_c0
    10  415    --   --  PF02435.11  Glyco_hydro_68    2.3e-70  Levansucrase/Invertase
    64  212    --   --  PF00251.15  Glyco_hydro_32N    0.0001  Glycosyl hydrolases family
                                                                 32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp16585_c0
     1   45    --   --  PF08244.7   Glyco_hydro_32C   9.4e-08  Glycosyl hydrolases family
                                                                 32 C terminal
## Pfam structure SIMILARITY for GENE: comp30971_c0
    76  109    --   --  PF08244.7   Glyco_hydro_32C   0.00011  Glycosyl hydrolases family
                                                                 32 C terminal


# Pfam INPUT sequence for gi|494111707|ref|WP_007052488.1| glycosyl hydrolase
   [Bifidobacterium longum]
# -- ALI -- -- ENV --
# BEGIN END  BEGIN END  PFAM_ID     SHORT_NAME        E-VALUE  DESCRIPTION OF TARGET
    35  345    35  346  PF00933     Glyco_hydro_3     3.2e-69  Glycosyl hydrolase family
                                                                 3 N terminal domain
   386  662   385  662  PF01915     Glyco_hydro_3_C   5.5e-42  Glycosyl hydrolase family
                                                                 3 C-terminal domain
   705  774   705  774  PF14310     Fn3-like          5.8e-18  Fibronectin type III-like
                                                                 domain
## Pfam structure SIMILARITY for GENE: comp17344_c0
     3  263    --   --  PF00933.16  Glyco_hydro_3     1.2e-34  Glycosyl hydrolase family
                                                                 3 N terminal domain
## Pfam structure SIMILARITY for GENE: comp5290_c0
    16  294    --   --  PF00933.16  Glyco_hydro_3     6.1e-30  Glycosyl hydrolase family
                                                                 3 N terminal domain
## Pfam structure SIMILARITY for GENE: comp5566_c0
    13  294    --   --  PF00933.16  Glyco_hydro_3     1.2e-29  Glycosyl hydrolase family
                                                                 3 N terminal domain
## Pfam structure SIMILARITY for GENE: comp6249_c0
    13  306    --   --  PF00933.16  Glyco_hydro_3     1.5e-50  Glycosyl hydrolase family
                                                                 3 N terminal domain


# Pfam INPUT sequence for gi|518557774|ref|WP_019727981.1| sucrose-6-phosphate hydrolase
   [Bifidobacterium breve]
# -- ALI -- -- ENV --
# BEGIN END  BEGIN END  PFAM_ID     SHORT_NAME        E-VALUE  DESCRIPTION OF TARGET
    34  332    34  335  PF00251     Glyco_hydro_32N   3.4e-97  Glycosyl hydrolases family
                                                                 32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp1218_c0
    14  326    --   --  PF00251.15  Glyco_hydro_32N   4.4e-88  Glycosyl hydrolases family
                                                                 32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp13476_c0
    32  125    --   --  PF00251.15  Glyco_hydro_32N   1.1e-39  Glycosyl hydrolases family
                                                                 32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp13900_c0
```

```
    76  277    --   --  PF00251.15  Glyco_hydro_32N   8.7e-27  Glycosyl hydrolases family
                                                               32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp13901_c0
     4   82    --   --  PF00251.15  Glyco_hydro_32N   4.4e-07  Glycosyl hydrolases family
                                                               32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp16517_c0
     5  171    --   --  PF00251.15  Glyco_hydro_32N   3.7e-19  Glycosyl hydrolases family
                                                               32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp17480_c0
    76  217    --   --  PF00251.15  Glyco_hydro_32N   3.4e-21  Glycosyl hydrolases family
                                                               32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp23569_c0
     7  153    --   --  PF00251.15  Glyco_hydro_32N   2.4e-21  Glycosyl hydrolases family
                                                               32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp23587_c0
     7  161    --   --  PF00251.15  Glyco_hydro_32N   4.4e-21  Glycosyl hydrolases family
                                                               32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp27727_c0
    10   96    --   --  PF00251.15  Glyco_hydro_32N     3e-05  Glycosyl hydrolases family
                                                               32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp28146_c0
    10   94    --   --  PF00251.15  Glyco_hydro_32N   7.8e-19  Glycosyl hydrolases family
                                                               32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp30085_c0
     5  117    --   --  PF00251.15  Glyco_hydro_32N   3.1e-11  Glycosyl hydrolases family
                                                               32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp6771_c0
    10  322    --   --  PF00251.15  Glyco_hydro_32N   1.5e-23  Glycosyl hydrolases family
                                                               32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp8872_c0
    21  334    --   --  PF00251.15  Glyco_hydro_32N   2.3e-23  Glycosyl hydrolases family
                                                               32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp30828_c0
     7  106    --   --  PF00251.15  Glyco_hydro_32N   2.2e-28  Glycosyl hydrolases family
                                                               32 N-terminal domain
    39  112    --   --  PF04616.9   Glyco_hydro_43    1.2e-05  Glycosyl hydrolases family 43
## Pfam structure SIMILARITY for GENE: comp2468_c0
    45  349    --   --  PF00251.15  Glyco_hydro_32N   8.9e-82  Glycosyl hydrolases family
                                                                 32 N-terminal domain
    63  240    --   --  PF04616.9   Glyco_hydro_43    2.4e-05  Glycosyl hydrolases family 43
   388  460    --   --  PF08244.7   Glyco_hydro_32C   1.1e-05  Glycosyl hydrolases family
                                                                 32 C terminal
## Pfam structure SIMILARITY for GENE: comp14192_c0
    33  343    --   --  PF00251.15  Glyco_hydro_32N   3.5e-97  Glycosyl hydrolases family
                                                                 32 N-terminal domain
   383  458    --   --  PF08244.7   Glyco_hydro_32C     1e-13  Glycosyl hydrolases family
                                                                 32 C terminal
## Pfam structure SIMILARITY for GENE: comp14228_c0
    33  343    --   --  PF00251.15  Glyco_hydro_32N  1.4e-100  Glycosyl hydrolases family
                                                                 32 N-terminal domain
   382  458    --   --  PF08244.7   Glyco_hydro_32C   5.7e-14  Glycosyl hydrolases family
                                                                 32 C terminal
## Pfam structure SIMILARITY for GENE: comp13535_c0
   186  624    --   --  PF02435.11  Glyco_hydro_68   9.8e-106  Levansucrase/Invertase
   311  428    --   --  PF00251.15  Glyco_hydro_32N   0.00019  Glycosyl hydrolases family
                                                                 32 N-terminal domain
## Pfam structure SIMILARITY for GENE: comp15259_c0
    10  415    --   --  PF02435.11  Glyco_hydro_68    2.3e-70  Levansucrase/Invertase
    64  212    --   --  PF00251.15  Glyco_hydro_32N    0.0001  Glycosyl hydrolases family
                                                                 32 N-terminal domain
```

```
            PFAM FOUND   SCORE  FREQ   R.F.  QUERY SEQUENCE
PF00251,PF04616,PF08244  0.33  =   1   4.0%  PF00251 Glyco_hydro_32N
PF00251,PF04616,PF08244  0.67  =   1   4.0%  PF00251 Glyco_hydro_32N,PF08244 Glyco_hydro_32C
        PF00251,PF04616  0.50  =   1   4.0%  PF00251 Glyco_hydro_32N,PF08244 Glyco_hydro_32C
        PF00251,PF08244  0.50  =   2   8.0%  PF00251 Glyco_hydro_32N
        PF00251,PF08244  1.00  =   2   8.0%  PF00251 Glyco_hydro_32N,PF08244 Glyco_hydro_32C
                PF00251  0.50  =  13  52.0%  PF00251 Glyco_hydro_32N,PF08244 Glyco_hydro_32C
                PF00251  1.00  =  13  52.0%  PF00251 Glyco_hydro_32N
                PF00933  0.33  =   4  16.0%  PF00933 Glyco_hydro_3,PF01915 Glyco_hydro_3_C,
                                               PF14310 Fn3-like
        PF02435,PF00251  0.50  =   2   8.0%  PF00251 Glyco_hydro_32N,PF08244 Glyco_hydro_32C
                PF08244  0.50  =   2   8.0%  PF00251 Glyco_hydro_32N,PF08244 Glyco_hydro_32C
-----------------------------------------------------
                         TOTAL  =  25
```

En los tres bloques se inicia con la arquitectura de consulta y en seguida se muestran las similitudes encontradas en el archivo del metagenoma. Cada una de las línas de similitudes que comienzan con ## representan una arquitectura de dominios en el archivo local. Estas arquitecturas presentan al menos una similitud de un 30 % con cada una de las arquitecturas de cada secuencia de entrada.

Al final del archivo se muestra una tabla resumen que muestra la arquitectura encontrada en el archivo, el puntaje $S_{SDA}$, la frecuencia, frecuencia relativa, la arquitectura de consulta, y el total de secuencias encontradas son contar duplicados.

A partir de estos resultados, el investigador podrá enfocarse en aquellos que son de su interés.

## 4.5. Comparación de SDA con otras herramientas

Con la finalidad de mostrar las bondades de uso de SDA, se realizó una tabla comparativa (Cuadro 4.13).

| | **T. promedio (hh:mm:ss)** | **Resultados gráficos** | **Personalización de estadísticos** | **Búsquedas locales** | **Múltiples secuencias** |
|---|---|---|---|---|---|
| **SDA** | 00:00:20 | × | × | × | × |
| **Pfam** | 06:41:28 | | × | | × |
| **CDART** | 00:05:53 | × | | | × |
| **WDAC** | 00:01:32 | × | | | × |
| **RADS** | N/A | × | | | × |

**Cuadro 4.13:** Tabla comparativa.

El tiempo promedio se refiere al tiempo que tomó obtener los resultados una vez que las secuencias están cargadas. SDA fue ejecutado en una computadora con procesador Intel Pentium CoreDuo a 0.83 GHz y memoria RAM de 4GB, también se ejecutó en una computadora con procesador AMD FX-6300 a 1.4 GHz con 16GB de memoria RAM obteniendo como tiempo promedio de ejecución 00:00:19. La personalización de estadísticos se refiere a si se permite cambiar libremente estos valores.

Múltiples secuencias es si el software permite hacer análisis de múltiples secuencias en una sola ejecución.

Para el caso de RADS/RAMPAGE y DDRO no se pudo medir los tiempos para obtención de resultados debido a problemas técnicos con su sitio web y documentación.

Como se mencionó en capítulos anteriores, el primer incoveniente con las herramientas actuales para identificación de dominios funcionales es la concurrencia de usuarios y la velocidad de red del usuario. Particularmente, el mayor tiempo esperado para obtener resultados fue cuando se usó el sitio de Pfam, con un promedio de espera de más de 6 horas.

RADS cuenta con una versión ejecutable que puede obtenerse desde el sitio http://www.mybio-software.com/rads-radscan-0-5-6-rampage-rapid-similarity-search-of-proteins-using-alignments.html, pero no cuenta con documentación de uso y el software arroja excepciones que no se sabe a qué se deben.

# Capítulo 5

# Discusión

Para todos los casos se realizaron comparaciones con las herramientas actuales, siendo Pfam la que genera los resultados más parecidos.

En el caso de las REasas en bacterias, los dominios encontrados por CDART, WDAC y Pfam son similares a los encontrados con SDA, salvo por los dominios empalmados que presenta SDA y que las demás herramientas no muestran pues conservan el dominio con los mejores estadísticos.

Los dominios encontrados son compartidos por todas las secuencias analizadas por lo que su estructura será la misma. SDA mostró ser capaz de mostrar resultados claros en enzimas con funciones y arquitecturas de dominios idénticas e incluso dominios que no muestran las demás herramientas que podrían ayudar a entender mejor al usuario cierta función que desempeña la proteína en cuestión.

En el segundo caso, para las REasas con sitios de reconocimiento similares, a pesar de que las enzimas presentaban sitios de reconocimiento idénticos o muy similares, los dominios encontrados no fueron los mismos. Para el caso de la secuencia 2 para AclI (*Streptomyces galilaeus*) y 6 para Psp1406I (*Helicobacter pylori*), cuyo sitio de reconocimiento y escición son idénticos, los dominios asociados (BTAD y Ribonuc L-PSP, respectivamente) no son los mismos. Del mismo modo, las estructuras de estos dominios no son similares, incluso, la estructura del dominios BTAD es más similar a aquella para el dominio RE Hind III, cuyo sitio de reconocimiento y corte es distinto al de AclI.

A diferencia del caso anterior, proteínas con funciones idénticas o muy similares, no siempre contarán con los mismos dominios asociados a ellas, es decir, son proteínas análogas, pues aunque la función es la misma, la composición no lo es. Proteínas como éstas, muestran como el proceso evolutivo puede tomar distintas vías para converger en una misma función. Esto también pudo observarse al obtener los resultados con SDA, por lo tanto, la herramienta desarrollada permite realizar búsquedas y análisis con mayo precisión, debido a que la búsqueda de términos semánticos como el nombre de la función, no es suficiente.

Para las enzimas con motivos Rossmann fold, todos los dominios están claramente diferenciados exceptuando los identificados en la secuencia 5 de la carbonil reductasa, en la que existe un empalme. Los dominios encontrados en las enzimas con motivos Rossmann fold son diferentes en todas las secuencias.

La búsqueda de dominios con las mismas secuencias en Pfam, arroja los mismos dominios, excepto que ignora el dominio *adh short*, debido en parte a su menor *e-value*.

De todas las enzimas con motivos Rossmann fold que fueron analizadas, en ninguna de ellas se observó una consevación a nivel de dominio como la existente a nivel estructural. Esta es una confirmación a la hipótesis de que la conformación de una proteína es más conservada que su secuencia. Además, ya que no hay documentación que indique que los dominios presentes en los motivos Rossmann fold son recurrente, los resultados obtenidos no permiten hacer conclusiones más detalladas sobre los dominios involucrados en la conformación característica de estos motivos.

Para el último caso, SDA ofrece ventajas sobre Pfam y otros recursos para la identificación de dominios en proteínas, y es que permite identificar dominios en una secuencia, y hacer una búsqueda local de arquitecturas similares en un archivo.

En cuanto a la búsqueda de las arquitecturas en el archivo de anotación del metagenoma del pulque, se encontraron coincidencias para las arquitecturas PF00251-PF08244 y PF00251.

Ya que los estadísticos del archivo de anotaciones están alejados de los usados con SDA para obtener resultados, se realizó otra ejecución pero ahora con un *e-value* de 0,1, una precisión de 0,7 y un *bit score* de 30. La tabla de similitudes es la siguiente:

```
            PFAM FOUND   SCORE   FREQ   R.F.  QUERY SEQUENCE
PF00251,PF04616,PF08244  0.67  =    1   4.0%  PF00251 Glyco_hydro_32N,PF08244 Glyco_hydro_32C
        PF00251,PF04616  0.50  =    1   4.0%  PF00251 Glyco_hydro_32N,PF08244 Glyco_hydro_32C
        PF00251,PF08244  1.00  =    2   8.0%  PF00251 Glyco_hydro_32N,PF08244 Glyco_hydro_32C
                PF00251  0.50  =   13  52.0%  PF00251 Glyco_hydro_32N,PF08244 Glyco_hydro_32C
                PF00933  0.33  =    4  16.0%  PF00933 Glyco_hydro_3,PF01915 Glyco_hydro_3_C,
                                                 PF14310 Fn3-like
        PF02435,PF00251  0.50  =    2   8.0%  PF00251 Glyco_hydro_32N,PF08244 Glyco_hydro_32C
                PF08244  0.50  =    2   8.0%  PF00251 Glyco_hydro_32N,PF08244 Glyco_hydro_32C
-----------------------------------------------
                         TOTAL =   25
```

Donde la cantidad de secuencias idénticas se redujo a solo una.

Este último caso describe de mejora manera las bondades del uso de SDA como herramienta en la identificación de dominios y el estudio de homólogos en proteínas, pues además de obtener la arquitectura de dominios de una o más secuencias de proteínas en una tabla resumen y de forma gráfica, permite realizar búsquedas de esas arquitecturas en una archivo con anotaciones.

# Conclusiones

El incremento en la información biológica generado por las nuevas ténicas de secuenciación propicia la creación de herramientas que permitan obtener distintos tipos de información que permitan a los investigadores obtener respuestas a cuestiones de su interés. La mayoría de las herramientas para la identifiación de dominios mencionadas en este trabajo tienen como principal ventaja que están disponibles para cualquier persona con acceso a Internet y sin la necesidad de instalar software adicional en su equipo.

Pero esas mismas ventajas pueden ser contraproducentes cuando, por ejemplo, no se cuenta con una buena conexión o incluso no se tiene conexión a Internet. Herramientas muy conocidas en la bioinformática, como Pfam, presenta inconvenientes cuando hay muchos usuarios que están haciendo uso de ella, lo que puede hacer esperar al usuario varias horas antes de obtener resultados.

La finalidad de la creación de SDA fue proveer alternativas a las herramientas ya existentes, en la que el usuario no dependa de recursos que en algún momento puede no tener. Las ventajas de SDA son que permite identificar dominios funcionales con modelos estocásticos fuertes en una cantidad relativamente corta de tiempo.

En las pruebas, SDA mostró ser más rápido al momento de obtener resultados analizando archivos con hasta ocho secuencias de aminoácidos.

La ventaja adicional de SDA al permitir realizar búsquedas de arquitecturas de dominios en archivos locales, facilita al usuario el análisis de sus datos reduciendo el tiempo que emplearía en hacer una búsqueda manual o en la construcción de un script para automatizarla.

Para los objetivos particulares, los resultados de SDA coincidieron con lo reportado en la literatura:

1. Hay proteínas con las mismas funciones, mismos dominios, y mismas estructuras, que posiblemente tienen un ancenstro evolutivo común del que “heredaron esas características”.

2. No todas las proteínas con la misma función o funciones similares presentan arquitecturas de dominios similares, incluso éstas pueden ser totalmente distintas, al igual que sus estructuras. Estas proteínas muestran casos de analogía, es

decir, de proteínas que aunque no tengan características estructurales similares, los procesos evolutivos las han conducido a obtener las mismas funciones.

3. Proteínas con las mismas estructuras superiores no significa que poseen las mismas funciones y por tanto, los mismos dominios. En el caso de los motivos Rossmann fold, es bien conocido que desempeñan funciones varias a pesar de tener la misma estructura.

4. Para el caso del metagenoma del pulque, SDA mostró se efectivo al momento de realizar búsquedas locales pues obtuvo resultados en cantidades relativamente cortas de tiempo.

En respuesta a la hipótesis planteada, de que es posible estudiar la arquitectura de dominios en proteínas para mejorar su anotación, con una estrategia que funcione en cualquier proteína y organismo, los resultados demuestran que a diferencia de las herramientas actuales que existen para el estudio de dominios funcionales, SDA demostró que se pueden obtener resultados robustos en menor tiempo y además, con funciones no encontradas en otras herramientas. El permitir al usuario mostrar todos los dominios encontrados en una secuencia junto con una representación gráfica, permite estudiar las arquitecturas en búsqueda de dominios en órdenes diferentes, con umbrales con mayor o menor sensibilidad, características que no están presentes en las demás herramientas con las que se suelen realizar estos estudios.

# Perspectivas

SDA ha mostrado buenos resultados, aunque existen áreas de mejora tales como la implementación de una base de datos para mostrar arquitecturas similares a las de la secuencia, así como la implementación de otro tipo de modelos para la identificación de dominios con la finalidad de dar resultados más concisos.

Actualmente se trabaja en la implementación de opciones para que el usuario pueda hacer búsquedas de arquitectura de dominios en genomas ya publicados.

# Bibliografía


[1] R. L. Schwartz, b. d foy, and T. Phoenix, *Learning Perl*, 6th ed. Sebastopol, CA. USA.: O'Reilly Media Inc., 2011.

[2] J. B. Hagen, "The origins of bioinformatics," *Nature Reviews Genetics*, vol. 1, no. 3, pp. 231–233, 2000. [Online]. Available: http://dx.doi.org/10.1038/35042090

[3] A. D. Baxevanis and B. F. Ouellette, *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, 2nd ed. New York, NY. USA: John Wiley & Sons, Inc., 2001.

[4] C. B. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 161, no. 4096, pp. 223–230, 1973. [Online]. Available: http://dx.doi.org/10.1126/science.181.4096.223

[5] H. G. S. C. International, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001. [Online]. Available: http://dx.doi.org/10.1038/35057062

[6] ——, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, no. 1, pp. 931–945, 2004. [Online]. Available: http://dx.doi.org/10.1038/nature03001

[7] N. Naidoo, Y. Pawitan, R. Soong, D. N. Cooper, and C.-S. Ku, "Human genetics and genomics a decade after the release of the draft sequence of the human genome," *Human Genomics*, vol. 5, no. 6, pp. 1–46, 2011. [Online]. Available: http://dx.doi.org/10.1186/1479-7364-5-6-577

[8] A. O. W. Stretton, "The first sequence: Fred sanger and insulin," *Genetics*, vol. 162, no. 2, pp. 527–532, 2002. [Online]. Available: http://www.genetics.org/content/162/2/527

[9] L. R. Croft, *Handbook of protein sequence analysis: a compilation of amino acid sequences of proteins with an introduction to the methodology*, 2nd ed. Chichester [Eng.]; New York: J. Wiley, 1980, first ed. published in

1973 under title: Handbook of protein sequences. [Online]. Available: http://trove.nla.gov.au/work/9341547

[10] B. o. Mathematical Biology, "Margaret oakley dayhoff 1925–1983," *Bulletin of Mathematical Biology*, vol. 46, no. 4, pp. 467–472, 1984. [Online]. Available: http://dx.doi.org/10.1007/BF02459497

[11] B. J. Strasser, "Collecting, comparing, and computing sequences: The making of margaret o. dayhoff's atlas of protein sequence and structure, 1954–1965," *Journal of the History of Biology*, vol. 43, no. 4, pp. 623–660, 2010. [Online]. Available: http://dx.doi.org/10.1007/s10739-009-9221-0

[12] M. Dayhoff, S. R.M., and O. B.C., *A model of evolutionary change in proteins*. Silver Spring:National Biomedical Research Foundation, 1978.

[13] S. Muff, "Lecture 4 - Evolutionary models and substitution matrices (PAM and BLOSUM)," 2011. [Online]. Available: http://www.math.uzh.ch/index.php?file&key1=19963

[14] T. Tatusova, S. Ciufo, B. Fedorov, K. O'Neill, and I. Tolstoy, "Refseq microbial genomes database: new representation and annotation strategy," *Nucleic Acids Research*, vol. 43, no. 7, p. 3872, 2015. [Online]. Available: http://nar.oxfordjournals.org/content/43/7/3872.short

[15] S. Goowdin, J. D. McPherson, and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nature Reviews Genetics*, vol. 17, no. 1, pp. 333–351, 2016. [Online]. Available: http://dx.doi.org/10.1038/nrg.2016.49

[16] J. M. Berg, J. L. Tymoczko, and L. Stryer, *Biochemistry*, 7th ed. New York, NY. USA: W. H. Freeman and Company, 2012.

[17] E. Feduchi-Canosa, I. Blasco-Castiñeyra, C. S. Romero-Magdalena, and E. Yañez-Conde, *Bioquímica. Conceptos Esenciales*. Madrid, España: Editorial Médica Panamericana, 2010.

[18] G. Karp, *Biología Celular y Molecular. Conceptos y Experimentos*, 5th ed. México, DF.: McGraw-Hill Interamericana Editores, 2009.

[19] R. Chang, *Química*, 7th ed. México, DF.: McGraw Hill Interamericana Editores, 2002.

[20] J. Luque-Cabrera and Á. Herráez-Sánchez, *Texto Ilustrado de Biología Molecular e Ingeniería Genética. Conceptos, Técnicas y Aplicaciones en Ciencias de la Salud*. Madrid, España: MMI Elsevier España, S.A., 2001.

[21] V. Pando-Robles and C. Ferreira-Batista, *Una Ventana al Quehacer Científico*. Cuernavaca, Morelos. México: Instituto de Biotecnología, UNAM, 2007. [Online]. Available: http://www.ibt.unam.mx/server/PRG.base?tit:-,tipo:doc, dir:libros_25aniv.html

[22] G. T. Solomons and C. B. Fryhle, *Organic Chemistry*, 10th ed. New York, NY. USA.: John Wiley & Sons, Inc., 2011.

[23] R. K. Murray, D. A. Bender, K. M. Botham, P. J. Kennelly, V. W. Rodwell, and P. A. Weil, *Harper. Bioquímica Ilustrada*, 29th ed. México, DF: McGraw Hill Interamericana Editores, 2010.

[24] D. L. Nelson and M. M. Cox, *Lehninger. Principles of Biochemistry*, 5th ed. New York, NY. USA: W.H. Freeman and Company, 2008.

[25] R. O. Esquivel, M. Molina-Espíritu, F. Salas, C. Soriano, C. Barrientos, J. S. Dehesa, and J. A. Dobado, *Decoding the Building Blocks of Life from the Perspective of Quantum Information*. InTech, 2013. [Online]. Available: http://www.intechopen.com/books/advances-in-quantum-mechanics/decoding-the-building-blocks-of-life-from-the-perspective-of-quantum-information

[26] G. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, "Stereochemistry of polypeptide chain configurations," *Journal of Molecular Biology*, vol. 7, no. 1, pp. 95–99, 1963. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022283663800236

[27] M. G. Rossman and A. Liljas, "Recognition of structural domains in globular proteins," *Journal of Molecular Biology*, vol. 85, no. 1, pp. 177–181, 1974. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0022283674901363

[28] I. Hanukoglu, "Proteopedia: Rossmann fold: A beta-alpha-beta fold at dinucleotide binding sites," *Biochemistry and Molecular Biology Education*, vol. 43, no. 3, pp. 206–209, 2015. [Online]. Available: http://dx.doi.org/10.1002/bmb.20849

[29] P. Laurino, Á. Tóth-Petróczy, R. Meana-Pañeda, W. Lin, D. G. Truhlar, and D. S. Tawfik, "An ancient fingerprint indicates the common ancestry of rossmann-fold enzymes utilizing different ribose-based cofactors," *PLoS Biology*, vol. 14, no. 3, pp. 1–23, 03 2016. [Online]. Available: http://dx.doi.org/10.1371/journal.pbio.1002396

[30] C. P. Ponting and R. R. Russell, "The natural history of protein domains," *Annual Review of Biophysics and Biomolecular Structure*, vol. 31, no. 1, pp. 45–71, 2002, pMID: 11988462. [Online]. Available: http://dx.doi.org/10.1146/annurev.biophys.31.082901.134314

[31] N. Song, R. Sedgewick, and D. Durand, "Domain architecture comparison for multidomain homology identification," *Journal of Computational Biology*, vol. 14, no. 4, pp. 496–516, 2007. [Online]. Available: http://dx.doi.org/10.1089/cmb.2007.A009

[32] C. Vogel, M. Bashton, N. D. Kerrison, C. Chothia, and S. A. Teichmann, "Structure, function and evolution of multidomain proteins," *Current Opinion in Structural Biology*, vol. 14, no. 2, pp. 208–216, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0959440X04000454

[33] C. Vogel, S. A. Teichmann, and J. Pereira-Leal, "The relationship between domain duplication and recombination," *Journal of Molecular Biology*, vol. 346, no. 1, pp. 355–365, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022283604015098

[34] M. Bashton and C. Chothia, "The geometry of domain combination in proteins," *Journal of Molecular Biology*, vol. 315, pp. 927–939, 2002. [Online]. Available: http://dx.doi.org/10.1006/jmbi.2001.5288

[35] S. K. Kummerfeld and S. A. Teichmann, "Protein domain organisation: adding order," *BMC Bioinformatics*, vol. 10, no. 1, pp. 1–11, 2009. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-10-39

[36] C. H. Wu, H. Huang, L.-S. L. Yeh, and W. C. Barker, "Protein family classification and functional annotation," *Computational Biology and Chemistry*, vol. 27, no. 1, pp. 37–47, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1476927102000981

[37] J. Xiong, *Essential Bioinformatics*. Cambridge, UK: Cambridge University Press, 2006.

[38] W.-K. Sung, *Algorithms in Bioinformatics a Practical Introduction*. Boca Raton, FL. USA: Taylor and Francis Group, LLC, 2010.

[39] I. Miklós, "Introduction to Algorithms in Bioinformatics," 2016. [Online]. Available: http://www.renyi.hu/~miklosi/AlgorithmsOfBioinformatics.pdf

[40] P. H. Sellers, "On the theory and computation of evolutionary distances," *SIAM Journal on Applied Mathematics*, vol. 26, no. 4, pp. 787–793, 1974. [Online]. Available: http://dx.doi.org/10.1137/0126070

[41] R. C. Edgar, "Muscle: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, 2004. [Online]. Available: http://nar.oxfordjournals.org/content/32/5/1792.abstract

[42] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson, and D. G. Higgins, "Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega," *Molecular Systems Biology*, vol. 7, no. 1, 2011. [Online]. Available: http://msb.embopress.org/content/7/1/539

[43] J. Pevsner, *Bioinformatics and Functional Genomics*, 2nd ed. Hoboken, NJ. USA.: John Wiley & Sons, Inc., 2009.

[44] W. R. Pearson and D. J. Lipman, "Improved tools for biological sequence comparison," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 85, no. 8, pp. 2444–2448, 1988.

[45] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic Local Alignment Search Tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1999. [Online]. Available: http://dx.doi.org/10.1016/S0022-2836(05)80360-2

[46] S. Henikoff and J. Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, no. 2, pp. 10 915–10 919, 1992. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/1438297

[47] D. T. Jones, W. R. Taylor, and J. M. Thornton, "The rapid generation of mutation data matrices from protein sequences," *Computer applications in the biosciences: CABIOS*, vol. 8, no. 3, pp. 275–282, 1992. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/8/3/275.abstract

[48] O. Gotoh, "An improved algorithm for matching biological sequences," *Journal of Molecular Biology*, vol. 162, no. 3, pp. 705 – 708, 1982. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0022283682903989

[49] P. Blunsom, "Hidden markov models," 2004. [Online]. Available: http://digital.cs.usu.edu/~cyan/CS7960/hmm-tutorial.pdf

[50] E. Fosler-Lussier, "Markov Models and Hidden Markov Models: A Brief Tutorial," 1998. [Online]. Available: http://web.cse.ohio-state.edu/~fosler/papers/tr-98-041.pdf

[51] G. Zitkovic, "Introduction to stochastic processes - lecture notes," 2010. [Online]. Available: https://www.ma.utexas.edu/users/gordanz/notes/introduction_to_stochastic_processes.pdf

[52] W. Ewens and G. Grant, *Statistical Methods in Bioinformatics: An Introduction*, 2nd ed. New York, NY. USA: Springer Science+Business Media, Inc., 2005.

[53] N. C. Jones and P. A. Pevzner, *An Introduction to Bioinformatics Algorithms*. Cambridge, MA. USA.: The MIT Press, 2004.

[54] S. R. Eddy and T. J. Wheeler, "HMMER's User Guide. Biological sequence analysis using profile hidden Markov models," 2015. [Online]. Available: http://eddylab.org/software/hmmer3/3.1b2/Userguide.pdf

[55] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, and A. Bateman, "The pfam protein families database: towards a more sustainable future," *Nucleic Acids Research*, vol. 44, no. D1, pp. D279–D285, 2016. [Online]. Available: http://nar.oxfordjournals.org/content/44/D1/D279.abstract

[56] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate, and M. Punta, "Pfam: the protein families database," *Nucleic Acids Research*, vol. 42, no. D1, pp. D222–D230, 2014. [Online]. Available: http://nar.oxfordjournals.org/content/42/D1/D222.abstract

[57] C. J. Sigrist, E. de Castro, L. Cerutti, B. A. Cuche, N. Hulo, A. Bridge, L. Bougueleret, and I. Xenarios, "New and continuing developments at prosite," *Nucleic Acids Research*, vol. 41, no. D1, pp. D344–D347, 2013. [Online]. Available: http://nar.oxfordjournals.org/content/41/D1/D344.abstract

[58] I. Letunic, T. Doerks, and P. Bork, "Smart: recent updates, new developments and status in 2015," *Nucleic Acids Research*, vol. 43, no. D1, pp. D257–D260, 2015. [Online]. Available: http://nar.oxfordjournals.org/content/43/D1/D257.abstract

[59] A. Marchler-Bauer, M. K. Derbyshire, N. R. Gonzales, S. Lu, F. Chitsaz, L. Y. Geer, R. C. Geer, J. He, M. Gwadz, D. I. Hurwitz, C. J. Lanczycki, F. Lu, G. H. Marchler, J. S. Song, N. Thanki, Z. Wang, R. A. Yamashita, D. Zhang, C. Zheng, and S. H. Bryant, "Cdd: Ncbi's conserved domain database," *Nucleic Acids Research*, vol. 43, no. D1, pp. D222–D226, 2015. [Online]. Available: http://nar.oxfordjournals.org/content/43/D1/D222.abstract

[60] A. Mitchell, H.-Y. Chang, L. Daugherty, M. Fraser, S. Hunter, R. Lopez, C. McAnulla, C. McMenamin, G. Nuka, S. Pesseat, A. Sangrador-Vegas, M. Scheremetjew, C. Rato, S.-Y. Yong, A. Bateman, M. Punta, T. K. Attwood, C. J. Sigrist, N. Redaschi, C. Rivoire, I. Xenarios, D. Kahn, D. Guyot, P. Bork, I. Letunic, J. Gough, M. Oates, D. Haft, H. Huang, D. A. Natale, C. H. Wu, C. Orengo, I. Sillitoe, H. Mi, P. D. Thomas, and R. D. Finn, "The interpro protein families database: the classification resource after 15 years," *Nucleic*

*Acids Research*, vol. 43, no. D1, pp. D213–D221, 2015. [Online]. Available: http://nar.oxfordjournals.org/content/43/D1/D213.abstract

[61] A. Andreeva, D. Howorth, J.-M. Chandonia, S. E. Brenner, T. J. Hubbard, C. Chothia, Murzin, and A. G., "Data growth and its impact on the SCOP database: new developments," *Nucleic Acids Research*, vol. 36, no. 1, pp. D419–D425, 2008. [Online]. Available: http://dx.doi.org/10.1093/nar/gkm993

[62] L. Y. Geer, M. Domrachev, D. J. Lipman, and S. H. Bryant, "Cdart: Protein homology by domain architecture," *Genome Research*, vol. 12, no. 10, pp. 1619–1623, 2002. [Online]. Available: http://genome.cshlp.org/content/12/10/1619.abstract

[63] B. Lee and D. Lee, "Protein comparison at the domain architecture level," *BMC Bioinformatics*, vol. 10, no. 15, pp. 1–9, 2009. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-10-S15-S5

[64] M. A. Messih, M. Chitale, V. B. Bajic, D. Kihara, and X. Gao, "Protein domain recurrence and order can enhance prediction of protein functions," *Bioinformatics*, vol. 28, no. 18, pp. i444–i450, 2012. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/28/18/i444.abstract

[65] N. Terrapon, J. Weiner, S. Grath, A. D. Moore, and E. Bornberg-Bauer, "Rapid similarity search of proteins using alignments of domain arrangements," *Bioinformatics*, vol. 30, no. 2, pp. 274–281, 2014. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/30/2/274.abstract

[66] R. A. Sayle and E. Milner-White, "RASMOL: biomolecular graphics for all," *Trends in Biochemical Sciences*, vol. 20, no. 9, pp. 374–376, 1995. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0968000400890805

[67] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000. [Online]. Available: http://dx.doi.org/10.1093/nar/28.1.235

[68] Y. Zheng, D. Cohen-Karni, D. Xu, H. G. Chin, G. Wilson, S. Pradhan, and R. J. Roberts, "A unique family of mrr-like modification-dependent restriction endonucleases," *Nucleic Acids Research*, vol. 38, no. 16, pp. 5527–5534, 2010. [Online]. Available: http://nar.oxfordjournals.org/content/38/16/5527.abstract

[69] R. J. Roberts, T. Vincze, J. Posfai, and D. Macelis, "Rebase—a database for dna restriction and modification: enzymes, genes and genomes," *Nucleic Acids Research*, vol. 43, no. D1, pp. D298–D299, 2015. [Online]. Available: http://nar.oxfordjournals.org/content/43/D1/D298.abstract

[70] C. A. Bottoms, P. E. Smith, and J. J. Tanner, "A structurally conserved water molecule in Rossmann dinucleotide-binding domains," *Protein Science*, vol. 11, no. 9, pp. 2125–2137, 2002. [Online]. Available: http://dx.doi.org/10.1110/ps.0213502

[71] J. Polaina, "Estructura, función e ingeniería molecular de enzimas implicadas en la digestión de carbohidratos," *Mensaje Bioquímico*, vol. 28, no. 1, pp. 61–76, 2004. [Online]. Available: http://bq.unam.mx/wikidep/uploads/MensajeBioquimico/Mensaje_Bioq04v28p061_Polaina_04.pdf

[72] B. Henrissat, "A classification of glycosyl hydrolases based on amino acid sequence similarities," *Biochemical Journal*, vol. 280, no. 2, pp. 309–316, 1991. [Online]. Available: http://www.biochemj.org/content/280/2/309

[73] J. Thompson, S. A. Robrish, S. Immel, F. W. Lichtenthaler, B. G. Hall, and A. Pikis, "Metabolism of sucrose and its five linkage-isomeric alpha-d-glucosyl-d-fructoses by klebsiella pneumoniae: Participation and properties of sucrose-6-phosphate hydrolase and phospho-alpha-glucosidase," *Journal of Biological Chemistry*, vol. 276, no. 40, pp. 37 415–37 425, 2001. [Online]. Available: http://www.jbc.org/content/276/40/37415.abstract

[74] W. Lammens, K. Le Roy, L. Schroeven, A. Van Laere, A. Rabijns, and W. Van den Ende, "Structural insights into glycoside hydrolase family 32 and 68 enzymes: functional implications," *Journal of Experimental Botany*, vol. 60, no. 3, pp. 727–740, 2009. [Online]. Available: http://jxb.oxfordjournals.org/content/60/3/727.abstract

[75] J. N. Varghese, M. Hrmova, and G. B. Fincher, "Three-dimensional structure of a barley beta-D-glucan exohydrolase, a family 3 glycosyl hydrolase," *Structure*, vol. 7, no. 2, pp. 179–190, 1999. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0969212699800240

[76] M. A. Schell, M. Karmirantzou, B. Snel, D. Vilanova, B. Berger, G. Pessi, M.-C. Zwahlen, F. Desiere, P. Bork, M. Delley, R. D. Pridmore, and F. Arigoni, "The genome sequence of bifidobacterium longum reflects its adaptation to the human gastrointestinal tract," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 22, pp. 14 422–14 427, 2002. [Online]. Available: http://www.pnas.org/content/99/22/14422.abstract

[77] B. Biavati, M. Vescovo, S. Torriani, and V. Bottazzi, "Bifidobacteria: history, ecology, physiology and applications," *Annals of Microbiology*, vol. 50, no. 1, pp. 117–131, 2000. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.583.7996&rep=rep1&type=pdf

[78] A. Silva, F. Barbosa, R. Duarte, L. Vieira, R. Arantes, and J. Nicoli, "Effect of bifidobacterium longum ingestion on experimental salmonellosis in mice," *Journal of Applied Microbiology*, vol. 97, no. 1, pp. 29–37, 2004. [Online]. Available: http://dx.doi.org/10.1111/j.1365-2672.2004.02265.x

[79] M. Bouvier, S. Meance, C. Bouley, J.-L. Berta, and J.-C. Grimaud, "Effects of consumption of a milk fermented by the probiotic strain bifidobacterium animalis dn-173 010 on colonic transit times in healthy humans," *Bioscience and Microflora*, vol. 20, no. 2, pp. 43–48, 2001. [Online]. Available: http://doi.org/10.12938/bifidus1996.20.43

[80] S. Meance, C. Cayuela, A. Raimondi, P. Turchet, C. Lucas, and J. michel Antoine, "Recent advances in the use of functional foods: Effects of the commercial fermented milk with bifidobacterium animalis strain dn-173 010 and yoghurt strains on gut transit time in the elderly," *Microbial Ecology in Health and Disease*, vol. 15, no. 1, pp. 15–22, 2003. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/08910600310015565

[81] R. Tanaka, H. Takayama, M. Morotomi, T. Kurishima, S. Ueyama, K. Matsumoto, A. Kuroda, and M. Mutai, "Effects of administration of tos and bifidobacterium breve 4006 on the human fecal flora," *Bifidobacteria and Microflora*, vol. 2, no. 1, pp. 17–24, 1983. [Online]. Available: http://dx.doi.org/10.12938/bifidus1982.2.1_17

[82] Y. Shimakawa, S. Matsubara, N. Yuki, M. Ikeda, and F. Ishikawa, "Evaluation of Bifidobacterium breve strain Yakult-fermented soymilk as a probiotic food," *International Journal of Food Microbiology*, vol. 81, no. 2, pp. 131–136, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168160502002246

# Apéndice A

# SDA user guide

## A.1. Before you start

Scan Domain Architecture (SDA) is an script intended to help you with the analysis of protein functional domain architecture and evolution.

## A.2. Feedback

Any comments, suggestions, bugs or feedback please contact to Ramon Flores, `ramon.flores.r@outlook.com`.

## A.3. Requirements

SDA requires:

- Perl v5.18.2 or later version.
- Perl SVG module for graphics.
- The Pfam-A.hmm database[1].
- The HMMER[2] suite for hmm profiles. SDA was developed and tested under HMMER v3.1b1, but current version v3.1b2 should work as well.
- For the GUI, you'll need Java 7 or later.

[1] `ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam29.0/`
[2] `http://hmmer.org/`

SDA makes some searches inside the PFAM database to annotate the input sequences. Here, we use the PFAM version 29 (latest, 30). After installing the HMMER suite please format your database for hmmscan searches.[3]

## A.4. File format and headers

To search a list of proteins with PFAM annotation in order to study their functional domain architecture you will need: (1) An input file (or PFAM list file) of the sequence(s) you want to find and must be amino acid sequences. This program doesn't work with nucleic acid sequences; (2) if you want to compare your sequence PFAM annotation to other proteins, an annotation list in the **trinotate** format can be used.

It's very important that your annotation file has the following requirements: the header for the Pfam column must begin with *pfam* (capitals or small letter, doesn't matter). If the file is a genome file, the header of the genes must be or contain *gene_id*, and for a transcriptome file must be or contain *transcript*; absence of some of these headers will result in a suddenly termination of this tool.

SDA allows you to search in an annotation file like Trinotate transcriptome results, for proteins with similar architectures or new combinations of the domains of PFAM domains of your interest.

SDA works with annotation files from genomes or transcriptomes as long as they have the required format.

## A.5. Usage

SDA creates up to three files depending on the results found. Please, verify next section for details.

The way you have to use SDA is as follows:

```
./SDA -OPTION <PARAMETER>...
```

The user's interest sequence might be a FASTA transcriptome sequence in a file, or a PFAM architecture in a file or as a list.

The PFAM list may have two forms, (1) as a file with one architecture per line and each PFAM joined by a comma without spaces; each PFAM can only have the letters *PF* and five digits. (2) As a parameter list, in which case this input is interpreted just as one architecture; you can list up to three PFAMS this way, each one joined

[3]To format the Pfam-A.hmm file, execute `hmmpress` *Pfam-A.hmm* in the terminal. This will create four binary files .h3{fimp}

by a comma and without spaces.

The all possible options and their description are the next:

| | |
|---|---|
| `-a\|--acc` <*VALUE*> | Defines the accuracy for each PFAM in the hmmscan summary table. It must be value between 0 and 1 (default, 0.85). |
| `-c\|--cpu` <*CPUs*> | Defines the number of CPUs used to create the HMM profile with `hmmscan`. |
| `-e\|--eval` <*VALUE*> | E-value threshold for `hmmscan` and result filtering (default, 1e-10). Must be positive with format input as an integer or in exponential notation. |
| `-f\|--fasta` <*FILE*> | The path for the FASTA file with amino acid sequence(s). |
| `-g` | Use it if your annotation file is a gene file. The expected annotation file input is a transcriptome. |
| `-h\|--help` | Displays this help menu. |
| `-l\|--list` <*FILE*> | Use it if your input is one or more PFAMs architecture(s) in a file. |
| `-o\|--output` <*PATH*> | Changes the output path directory (default, parent working directoty). |
| `-p\|--pfam` <*FILE*> | Indicates the location of your Pfam-A.hmm file. |
| `-t\|--trin` <*FILE*> | Tab-delimited file with PFAM annotation for the genome/transcriptome file in Trinotate format. |
| `-v\|--version` | Shows the SDA version. |

## A.6. Output Files

SDA generates up to three different files in directory called *SDA.Results* in the output directory (∼/ as default).

If you feed SDA with a FASTA file, inside the results directory, SDA will create another directory with the name of your FASTA file and inside, you'll find the results. That means you can work with different FASTA files and you'll find your results in each different directory.

If you feed SDA with a PFAM architectures file, the result will be saved inside the *SDA.Results* directory if there is at least one coincidence or similarity between the PFAMs input and the annotation file.

The output files and their content is described below:

**.table** Contains the hmmscan results for the `hmmscan` fasta file. Please, check the `hmmscan` documentation for further details.

**.pdf** This file contains a graphical representation of your result for each sequence in the FASTA file. It draws the domains found for the sequence(s) and it's PFAM name.

**.out** It's created if SDA finds at least one coincidence or similarity between each FASTA sequence (or PFAM input) and the annotation file. Contains the gene_id or transcript, location, e-value, and description of the match. At the end of this file you'll find a summary with all the PFAM architectures found, similarity score and it's relative frequency.

## A.7. Similarity score

The similarity of the PFAMs for the `.out` file depends on the PFAM cardinality of each sequence in the FASTA file or PFAM list, and the number of matches. This way, the similarity scores is defined as

$$S_{SDA} = \frac{n}{\max\{l_1, l_2\}} \tag{A.1}$$

where $n$ is the amount of domains shared by the both of sequences, and $l_1, l_2$ are the total amount of domains found for each sequence.

## A.8. Working with SDA

There are many ways to use SDA, we shown some ways to use it.

### A.8.1. Identifying domains architectures with SDA

The first use of SDA is to identify domains in a sequence of amino acids. The simplest way to do this is typing

`$ ./SDA -p` *<Pfam-A path>* `-f` *<fastaFile path>*

SDA shown the results of processing the file and at the end, it will indicate you where you will find the results.

SDA will create the files `.table` and `.pdf` for results (figure A.1).

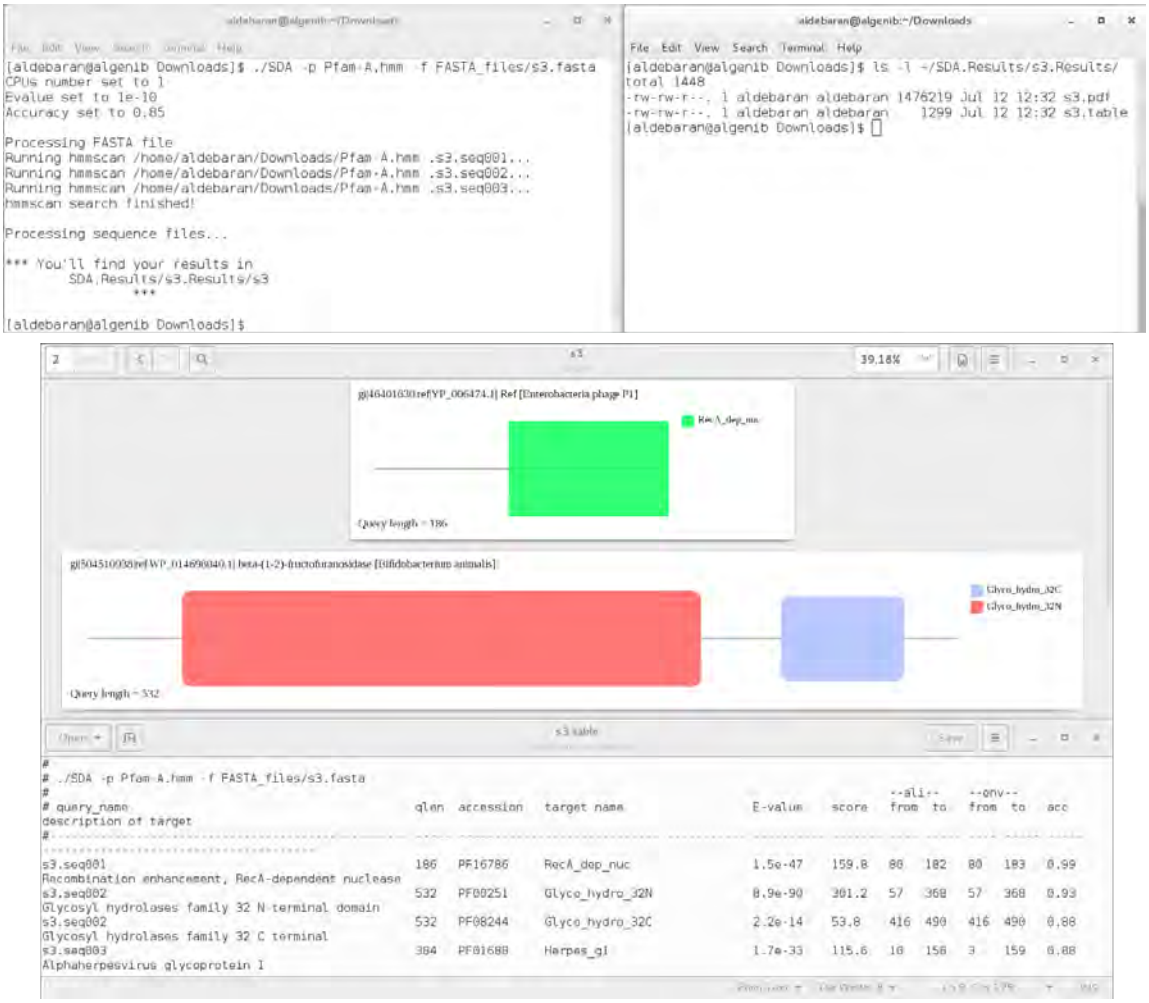



**Figura A.1:** Results from SDA. Up, example of run. Down, .table and .pdf files results.

Additionaly, you can indicate the number of CPUs used to create the `hmmscan` profiles, the e-value treshold, and the minimum accuracy of the results adding

`-c` <*CPUs*> `-a` <*accuracy*> `-e` <*e-value*>

## A.8.2. SDA with a FASTA file and an annotation file

For searching domain architectures in a trinotate annotation file, you must add the option `-t` with the path to the annotation file, and the option `-g` if the annotation

is a genome file (default is a transcriptome). Remember that the column names in the annotation file must have a format (section 4).

This way, the command will be

`$ ./SDA -p` <*Pfam-A path*> `-f` <*fastaFile path*> `-t` <*annotationFile path*> `-g`

And again, you can specify the number of processors and accuracy of results. The example run and results are shown in figure A.2.

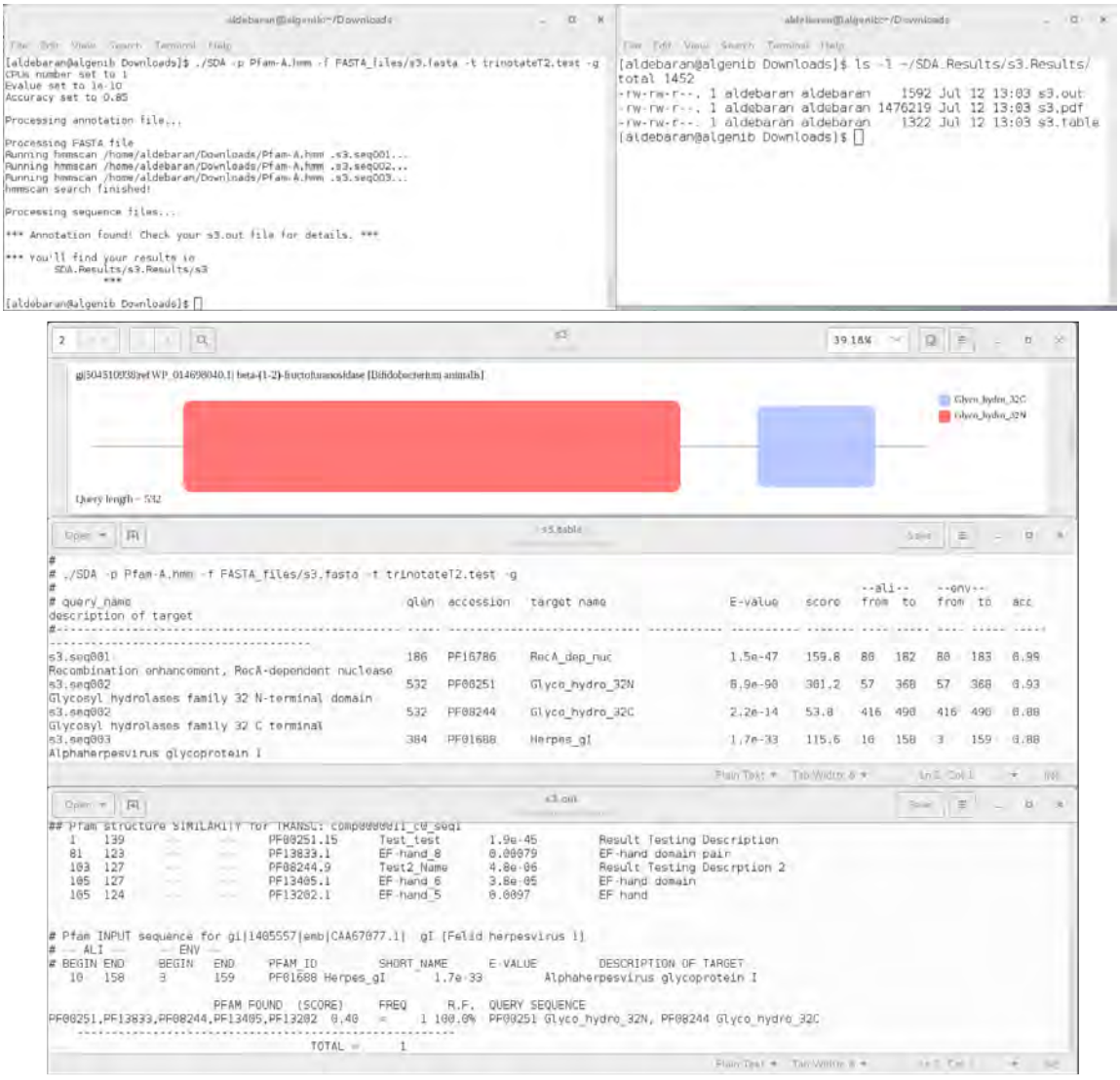



**Figura A.2:** Results from SDA with annotation file. Up, example of run. Down, .table, .pdf and .out files results.

The `.out` file contains a table with information of the query domain and the domain(s) found. At the end, there is a resume with the domain architectures found.

### A.8.3. SDA with a PFAM architectures file and an annotation file

SDA can also be used as a PFAM architectures searcher. If you have an annotation file and you are looking for some domains architectures, you can put the architectures in a file `.txt` and look for them in the annotation file. To do this, type

`$ ./SDA -l` *<Pfam Architectures path>* `-t` *<annotationFile path>* `-g`

For this usage, the values of accuray, e-value and CPUs are nonsense because SDA will just look for architectures and will not build the HMM profile.

You will get just a file in the home directory called as the input file with extension `.out`. This file will contain the similarities found (if there are) and the distribution resume. If there are non results, no one file will be created.

Figure A.3 shows an ejecution example and the results.

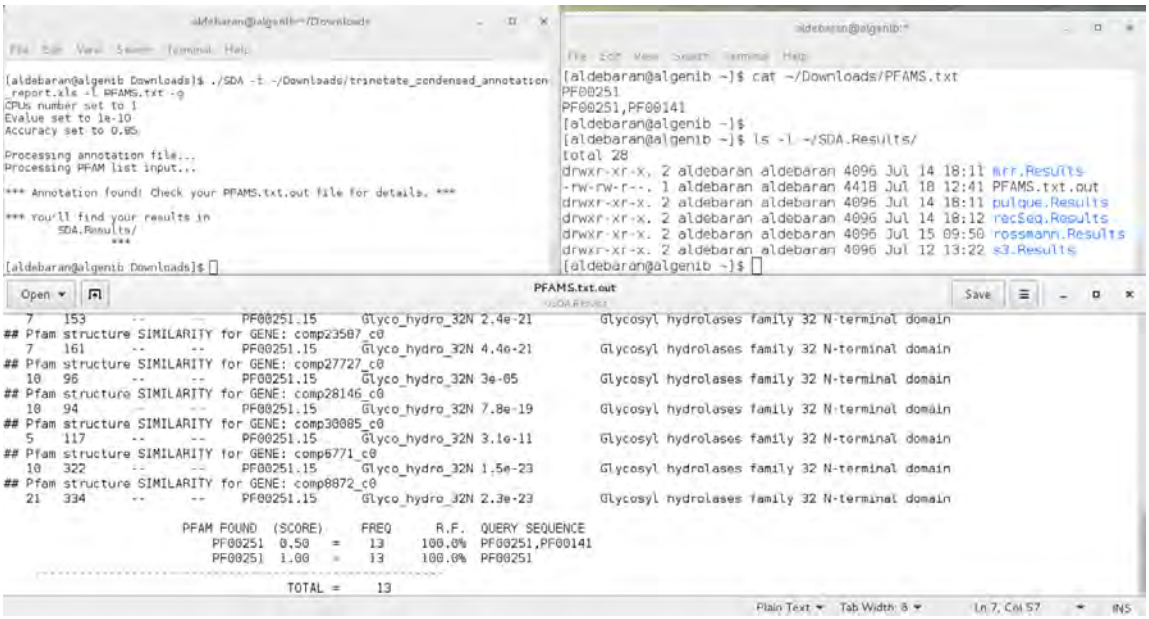



**Figura A.3:** Results from SDA with PFAM architecture file and annotation file. Up left, example of run. Up right, PFAM file content and result file. Down, .out file content.

## A.9. Using the GUI

There is a GUI for the users that dislike using shell, it is call `SDA.run` and was created with Java.

To execute it just open terminal in the folder where you have the files `SDA.run` and `SDA` (both of them must be in the same location) and type

```
$ ./SDA.run
```

The GUI is easy to use, you just have to select the files in your computer and click the button **start** to get results (figure A.4).

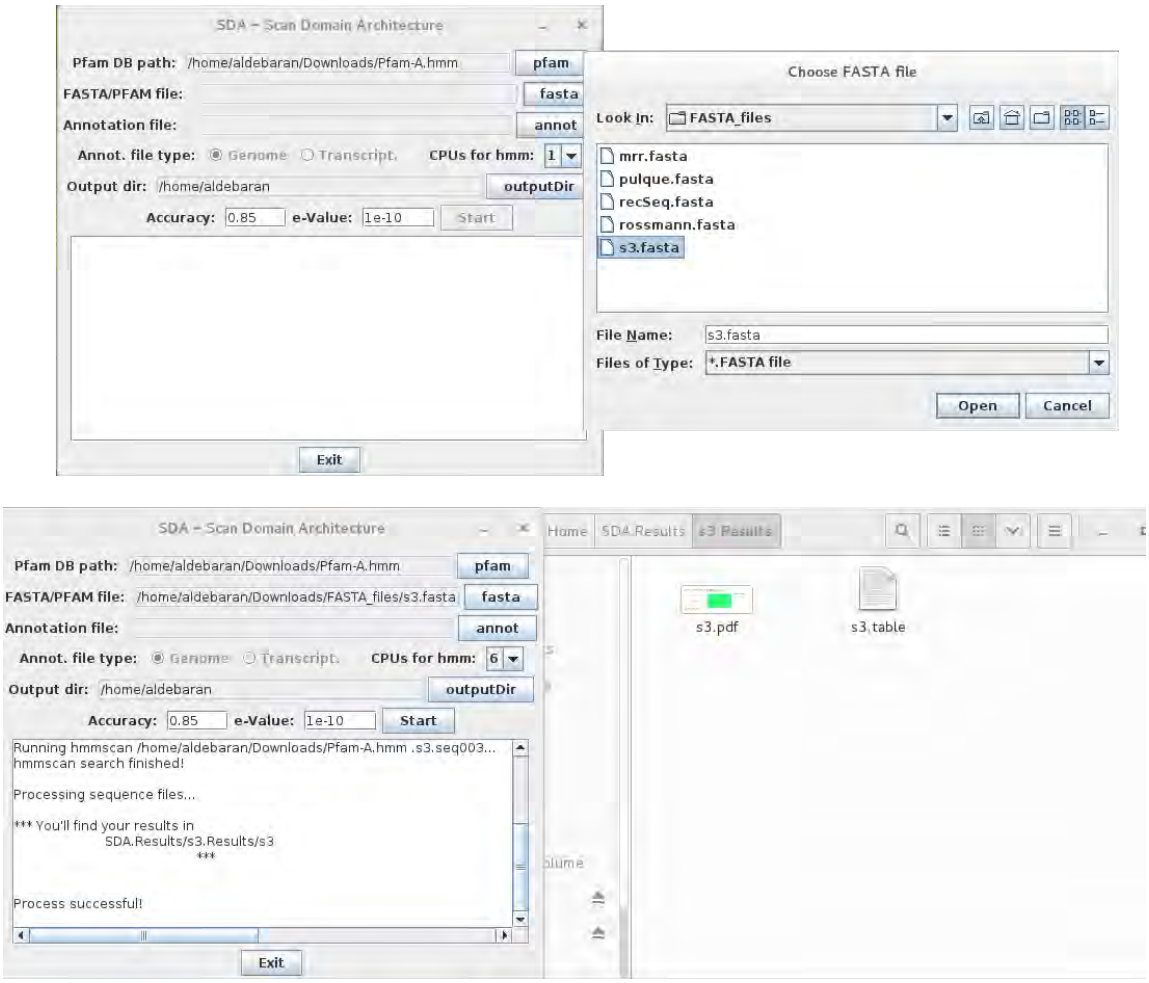



**Figura A.4:** Example of use of SDA GUI.

If you want to use SDA GUI as a Pfam architecture searcher, select the PFAM file with the FASTA button, and the annotation file with the annot button.

When the program finishes, SDA automatically will open the file with the results.