DESARROLLO DE APLICATIVO WEB PARA LA PREDICCIÓN DE ESTRUCTURAS SECUNDARIAS Y TERCIARIAS A PARTIR DE SU SECUENCIA DE AMINOÁCIDOS

OSCAR JULIAN CRUZ SALAZAR
PABLO ANDRES HENAO OLIVEROS

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE INGENIERÍAS
INGENÍERIA ELECTRÓNICA
Pereira- Risaralda
2015

DESARROLLO DE APLICATIVO WEB PARA LA PREDICCIÓN DE ESTRUCTURAS SECUNDARIAS Y TERCIARIAS A PARTIR DE SU SECUENCIA DE AMINOÁCIDOS

OSCAR JULIAN CRUZ SALAZAR
PABLO ANDRES HENAO OLIVEROS

Proyecto de grado para optar al título de:
Ingeniero Electrónico.

Director
Mauricio Alexander Álvarez López, PhD.

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE INGENIERÍAS
INGENÍERIA ELECTRÓNICA
Pereira- Risaralda
2015

AGRADECIMIENTOS

Primero que todo dedico el presente trabajo a Dios y a mis padres Gilma Oliveros y Carlos Alberto Henao que siempre estuvieron ahí en el momento que los necesitaba, gracias a ellos por acompañarme y darme la oportunidad de ser la persona que soy hoy en día. También doy gracias a todas las personas que estuvieron en esta época de mi vida como amigos, compañeros de clase, profesores, hermanos y compañero de proyecto Oscar Julián Cruz que me ayudaron a formarme como todo un profesional y como una persona. También gracias por el apoyo al director del trabajo de grado PhD Mauricio A. Álvarez, quien con su apoyo y conocimiento nos guio en la realización de este trabajo de grado.

Mil gracias a todas estar personas.

Pablo Andrés Henao Oliveros

Quiero aprovechar este espacio para expresar mi inmensa gratitud a mis padres, principales gestores de mi formación como ingeniero electrónico, a mi madre Liliana Salazar Gómez que fundo en mi todos los pensamientos emprendedores, de esfuerzo y fortaleza durante un camino complejo y largo como fue este que estamos por culminar, a mi padre Oscar Cruz Ramírez que fue un gran ejemplo de perseverancia y que me apoyo durante toda mi carrera universitaria. Quiero agradecer de igual manera a mis amigos y familia, a mi amigo Juan Pablo Holguín que alguna vez me aconsejo y mi compañero Pablo Andrés que materializo conmigo este sueño, de igual manera quiero agradecer a mis compañeros de clase que en su medida aportaron en mi conocimiento y nuevas formas de ver la vida y por ultimo pero no menos importante a mi maestro y ejemplo de vida el director del trabajo PhD Mauricio A. Álvarez que nunca dejo de guiarnos en este proceso.

Oscar Julián Cruz Salazar

CONTENIDO

Pág.

LISTA DE FIGURAS

1. INTRODUCCIÓN

La predicción de la estructura tridimensional de una proteína cuando sólo se conoce su secuencia de aminoácidos ha sido un problema de gran interés desde tiempo atrás. Los enfoques han variado desde el método ab-initio que se basa exclusivamente en principios físicos y químicos, al método de homología que se basa principalmente en la información disponible en las bases de datos de secuencias y estructuras. Los métodos antes mencionados se han convertido en modelos más precisos y su rango de aplicación se ha estado incrementando. Por el lado del ab-initio se debe a los grandes recursos computacionales los cuales permiten la implementación de procesos estocásticos más robustos y al avance en la comprensión de la base química de la proteína para un correcto plegado [1]. Por otro lado el modelo por homología tiene un avance significativo debido a la gran cantidad de secuencias y estructuras que se han obtenido y almacenado en bases de datos en los últimos años [2].

Para obtener resultados en la predicción de una proteína mediante el modelado por homología se requiere de múltiples procesos a realizar, tales como la determinación de la proteína más parecida en una base de datos, la alineación de la proteína de estudio con su proteína más cercana y la creación de un modelo tridimensional adecuado para la secuencia de aminoácidos [3]. En los procesos de predicción de estructuras de proteínas mediante el método ab initio se puede contar con un algoritmo diferente de acuerdo a cada servidor de pruebas encontrado en la web, esto termina en resultados muy diferentes para cada predicción realizada.

Los procesos de determinación de estructuras de proteínas en Colombia se ven limitados debido a la escases de recursos (económicos, académicos y de infraestructura) y a la falta de elementos que permitan obtenerlas experimentalmente. Ante las dificultades presentadas para obtener la

estructura de proteínas mediante métodos experimentales, surgen los métodos de predicción de estructuras de proteínas mediante software. Los métodos de predicción de estructuras de proteínas mediante software presentan según los resultados del Critical Assessment of Techniques for Protein Structure Prediction (CASP) [4], altos índices de acierto de acuerdo a los criterios de evaluación de estos. En Colombia existe una gran cantidad de materia biológica que requiere ser estudiada a nivel de la estructura de sus proteínas.

Existen diferentes frutas como el lulo que se producen en la región andina que necesitan ser mejoradas genéticamente, a su vez mediante el conocimiento de estructuras de proteínas se pueden generar enzimas con las cuales se pueden producir medicamentos, entre otras aplicaciones. Como ejemplo del problema, el lulo como otras frutas presenta un fenómeno denominado pardeamiento, este fenómeno consiste en la oxidación de parte de la fruta disminuyendo su calidad. El pardeamiento podría ser eliminado gracias al estudio y modificación de la proteína que genera este fenómeno.

En la actualidad existen programas y herramientas que facilitan la predicción de estructuras de proteínas así como la obtención de otro tipo de datos asociados a estas, sin embargo, los procesos necesarios para lograr resultados son complejos y requieren de la utilización de múltiples herramientas para un único fin (obtener la estructura 2D o 3D de la proteína). En Colombia los procesos de predicción de proteínas mediante software no son muy utilizados por la comunidad académica.

La creación de algoritmos de computación que reúnan información presente en la web y utilicen diferentes procesos para obtener la predicción de estructuras de proteínas, permite la recopilación de gran cantidad de datos. Reunir información y funcionalidades que solucionen requerimientos a partir de la información presente en la web requiere de la automatización de procesos

mediante algoritmos. La automatización de procesos consiste en un programa que ejecuta ciertos pasos para realizar un proceso específico y que se acopla a la necesidad determinada para realizar sistemáticamente su función. Mediante la automatización de procesos se podría obtener información de diferentes servidores de predicción de estructuras de proteínas para su recopilación, organización y comparación, de acuerdo a una secuencia de aminoácidos perteneciente a una proteína de estudio.

En Colombia no se ha desarrollado una herramienta que reúna información de diferentes aplicativos web de predicción de estructuras de proteínas. La reunión de diferentes métodos en un solo espacio permitiría facilitar los procesos de obtención y reunir gran cantidad de información necesaria para la investigación de las mismas, además de la reducción de costos en el proceso de predicción.

Un aplicativo de predicción de estructuras de proteínas es de suma importancia para el área de la bioinformática en particular para el centro de biología molecular y biotecnología (CENBIOTEP) perteneciente a la universidad tecnológica de Pereira debido su aplicabilidad en la predicción terciaria de la proteína, la cual puede ser utilizada para obtener la información suficiente para llevar acabo el mejoramiento genético en una proteína.

2. OBJETIVOS

En este capítulo se describe los alcances del proyecto.

a) OBJETIVO GENERAL

Desarrollar una aplicación web que permita la predicción de estructuras de proteínas secundarias y terciarias a partir de su secuencia de aminoácidos, que reúna múltiples algoritmos de predicción de estructuras de proteínas y que permita la visualización de datos para el análisis de las mismas.

b) OBJETIVOS ESPECIFICOS

i. Desarrollar algoritmos de programación que permitan navegar por diferentes servidores web de predicción de estructuras de proteínas y obtener los resultados de cada servidor en la predicción de estructuras de proteínas a partir de una cadena de aminoácidos.

ii. Desarrollar una interfaz web que permita, a través de una tubería de software, la predicción de la estructura 2D y 3D de una proteína a partir de su secuencia de aminoácidos mediante el método Ab Initio y mediante el método Homología.

iii. Validar los resultados de la aplicación web a partir de proteínas cuyas estructuras se encuentran ya determinadas en la base de datos del NCBI (National Center for Biotechnology Information) y PDB (Protein Data Bank).

3. METODOS PARA LA OBTENCIÓN DE ESTRUCTURAS DE PROTEÍNAS

En este capítulo se explicarán brevemente algunos conceptos acerca de modelos de estructuras de proteínas, posteriormente también se explicarán los métodos utilizados en la obtención de estructuras de proteínas, estos métodos permiten obtener un modelado de la estructura de la proteína a partir de su secuencia de aminoácidos. Los modelos entregados por estos métodos permiten ser utilizados para el mejoramiento genético de una gran cantidad de materia biológica, a su vez se pueden generar nuevas enzimas con las cuales se produce medicamentos entre otras aplicaciones.

3.1. Enlaces peptídicos.

Este tipo de enlace está conformado por un enlace entre un grupo **amino** de un aminoácido y el grupo **carboxilo** de otro aminoácido. [5]

3.2. Estructura primaria de proteínas.

Es la forma más básica de las proteínas. Este tipo de estructura está determinada por la secuencia de aminoácidos o por el número de aminoácidos presentes y por el orden en que están enlazados por medio de enlaces peptídicos. [6]

3.3. Estructura secundaria de proteínas.

La estructura secundaria de una proteína es el plegamiento que la cadena polipeptídica adopta debido a la formación de puentes de hidrógeno entre los átomos que conforman el enlace peptídico. Los puentes de hidrógeno se establecen cuando se comparte un protón entre dos moléculas, esto a su vez formando un enlace débil.

La predicción de la estructura secundaria son un conjunto de procesos bioinformáticos, su principal objetivo es la predicción de la estructura secundaria de proteínas y ácidos nucleicos, esto a partir de su secuencia de aminoácidos. La mayoría de los métodos utilizados para la construcción del modelo, se basan principalmente en el uso de redes neuronales y la comparación con modelos ya determinados que se encuentran en una base de datos. [7]

3.4. Estructura terciaria de proteínas.

La estructura terciaria de una proteína describe el plegamiento de los elementos de la estructura secundaria y especifica las posiciones de cada átomo en la proteína, incluidos los de sus cadenas laterales. [8]

3.5. Modelado por homología.

La predicción por homología o comparación consiste como su nombre lo indica en la comparación de las estructuras de proteínas obtenidas por algoritmos computacionales con estructuras de proteínas similares y ya conocidas. Estas cuentan con un modelo determinado en una base de datos de proteínas determinada, de esta manera al analizar un modelo por predicción computacional se compara con la obtenida mediante métodos de laboratorio, lo cual brinda el porcentaje de acierto de la predicción [9].

En la figura 1 se observa el diagrama de procesos general para la predicción de la estructura tridimensional de una proteína a partir de su secuencia de aminoácidos usando el modelo por homología.

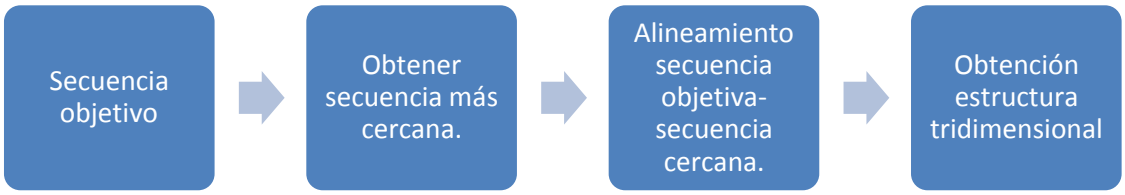


Figura 1. Pasos para la obtención del modelado por homología.

En la figura 1 se visualizan los diferentes pasos que se deben tener en cuenta para obtener el modelo estructural tridimensional de la secuencia de aminoácidos objetivo mediante el modelado por homología. Para la obtención de la secuencia más cercana se pueden llegar a utilizar diferentes servidores como son Psiblat y PDB. Lo que estos servidores realizan es buscar en sus respectivas bases de datos, secuencias de aminoácidos que contengan una mayor similitud entre sus componentes con respecto a la secuencia de aminoácidos objetivo. Posteriormente habiendo obtenido la secuencia más cercana se procede a realizar un alineamiento con servidores como Clustal-Omega, el alineamiento consiste en tomar las proteínas (secuencia objetivo y secuencia más cercana), y observar que caracteres son idénticos. En la etapa de metodología se explicaran estos métodos un poco más a fondo para su mayor entendimiento.

3.6. Modelado ab-Initio

La predicción de proteínas por el método AB-Initio tiene como objetivo principal la construcción de modelos de estructuras de proteínas a partir de su secuencia de aminoácidos desde un punto cero, basándose principalmente en principios físicos, a diferencia del modelado por homología el cual obtiene la estructura de proteínas a partir de una ya conocida. Los procedimientos posibles que se pueden realizar mediante el modelado ab-initio se basan en procesos estocásticos, redes neuronales, mapas de contacto y máquinas de soporte vectorial, estos procedimientos consisten en diferentes tipos de

algoritmos que por su estructura y resultados arrojan porcentajes de acierto distintos [10].

En la figura 2 se puede observar el diagrama de procesos general para la predicción de estructuras de proteínas mediante algoritmos de programación usando el modelado Ab initio o Novo. Los métodos de predicción por ab-initio pueden presentar una gran cantidad de algoritmos diferentes de acuerdo al autor que desarrolle el programa de predicción lo cual a su vez hace que se vea afectado de forma notable el porcentaje de acierto de las aplicaciones que funcionan bajo este método.

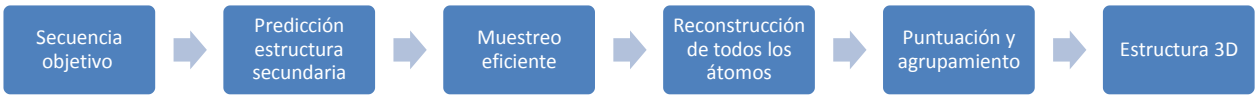


Figura 2. Pasos para la obtención del modelado AB-Initio.

3.7. Aplicación WEB.

Una aplicación web está constituida por varias páginas web que interactúan entre sí, utilizando los recursos en un servidor. Se utilizan en algunos casos para consultar, modificar o insertar, por medio del servidor, la información de las bases de datos.

Para el desarrollo de aplicaciones web comúnmente se utiliza el lenguaje PHP. Este lenguaje de programación tiene su aparición en el año 1994, desde entonces ha experimentado un gran crecimiento y acogida en el mundo, debido a las características que lo definen como son la potencia, versatilidad, robustez y modularidad [11]. Los programas escritos en PHP (Hypertext Pre-processor), son embebidos directamente en el código HTML (Hiper text Markup Language), ejecutado e interpretado por un servidor web antes de transferir al

usuario un resultado en lenguaje HTML puro. Además este lenguaje es de fácil aprendizaje por su flexibilidad y gran similitud en sintaxis a diferentes lenguajes.

Una de las características más destacadas del lenguaje PHP, es la fácil conectividad con sistemas gestores de bases de datos, como MySQL, lo cual ha generado la gran utilización para la creación de páginas dinámicas, no solo personales sino también portales empresariales [11].

3.8. Tubería de software o arquitectura pipeline.

En informática la tubería de software consiste en una cadena de pasos o procesos conectados de forma que la salida de cada proceso es la entrada del siguiente proceso. [12]

3.9. PhantomJS.

Es un framework de programación que permite parsear [1] código JavaScript como código cmd (Consola de Comandos) de Windows. Por lo cual permite manejar procesos del computador además está diseñado para hacer webdriving (manejo de web automático) y manejar páginas web de manera automatizada. [13]

[1] Parsear: Transformación de un tipo de variable o código completo en un tipo de variable diferente.

4. METODOLOGÍA

En este capítulo se describe el diseño metodológico llevado a cabo durante el desarrollo del proyecto. El programa utilizado para el desarrollo de los algoritmos para la predicción de estructuras de proteínas fue Sublime Text 2, el cual tiene una licencia de uso libre. Para el desarrollo de los algoritmos se utilizaron lenguajes de programación tales como: Html5, Css3, JavaScript, PHP y los frameworks JQUERY y PhantomJS.

En la aplicación web diseñada, todos los algoritmos desarrollados están generados en una arquitectura Pipeline o tubería de software. De tal manera que las diferentes metodologías utilizadas y las cuales se describirán posteriormente consisten en enviar la información desde la interfaz web a un servidor PHP. Este servidor evalúa si la proteína a predecir existe en una base de datos MYSQL y este retorna la información almacenada si la proteína existe o devolverá una bandera de no existencia si la proteína no es encontrada en esta base de datos. Si la proteína no existe en la base de datos, por medio de la interfaz se realiza un llamado al servidor PHP para la respectiva predicción. También el servidor PHP funciona como un intermediario entre la interfaz WEB y los algoritmos de procesamiento de PHANTOMJS debido a que estos últimos algoritmos deben ejecutarse sobre el sistema operativo. Este servidor se encarga de la ejecución de los algoritmos de procesamiento desarrollados en el Framework PhantomJS, la obtención de resultados de estos algoritmos y retorna estos resultados a la interfaz web. Todos los resultados obtenidos por medio de la interfaz y los cuales no se encuentren en la base de datos serán almacenados en esta para su posterior verificación y reutilización. Lo anterior mencionado se puede observar en la figura 3.

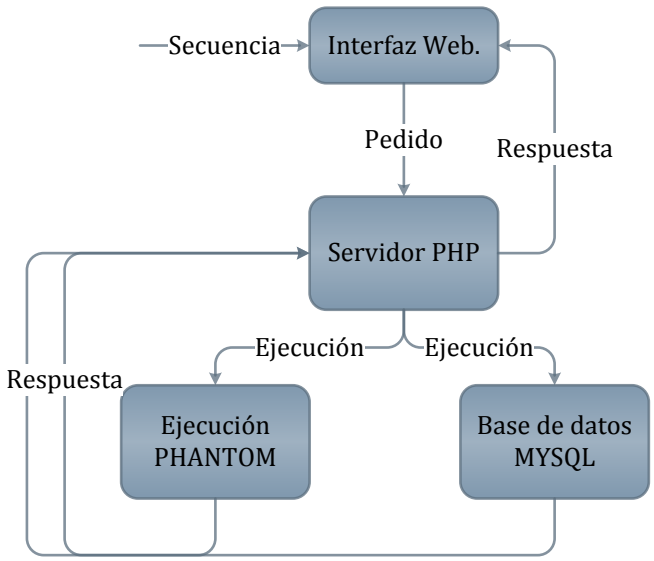


Figura 3. Diagrama de flujo funcionamiento general de la aplicación.

La herramienta web cuenta con tres tipos de modelado como son:

- Predicción Ab-Initio 3D.
- Predicción Ab- Initio 2D.
- Predicción homología 3D.

Como primer paso para el desarrollo del proyecto, se desarrolló una interfaz web, por medio de la cual el usuario pudiera ingresar una secuencia de aminoácidos en formato FASTA a predecir. Además se desarrolló una base de datos en MYSQL donde se guardan los datos de las proteínas ya obtenidas anteriormente por la aplicación, esto se realizó con el fin de disminuir los tiempos de predicción en dichas proteínas. Con respecto al formato de la secuencia se realizan diferentes validaciones para obtener un formato específico de esta. Es así que utilizando tecnología AJAX se envía una petición al servidor PHP, en este servidor se evalúa el tipo de llamado realizado y este evalúa en una base de datos MYSQL si la proteína ya existe o ha sido obtenida anteriormente, si la proteína ingresada existe se toman los datos de la información de la predicción deseada desde la base de datos.

Continuando con el proceso para el desarrollo de esta aplicación, si la proteína no existe en la base de datos se notifica a la interfaz web y se procede a realizar el proceso de predicción mediante el uso del método de homología. En la figura 4 se observa el diagrama de flujo para validar la existencia de la secuencia de proteína en la base de datos.

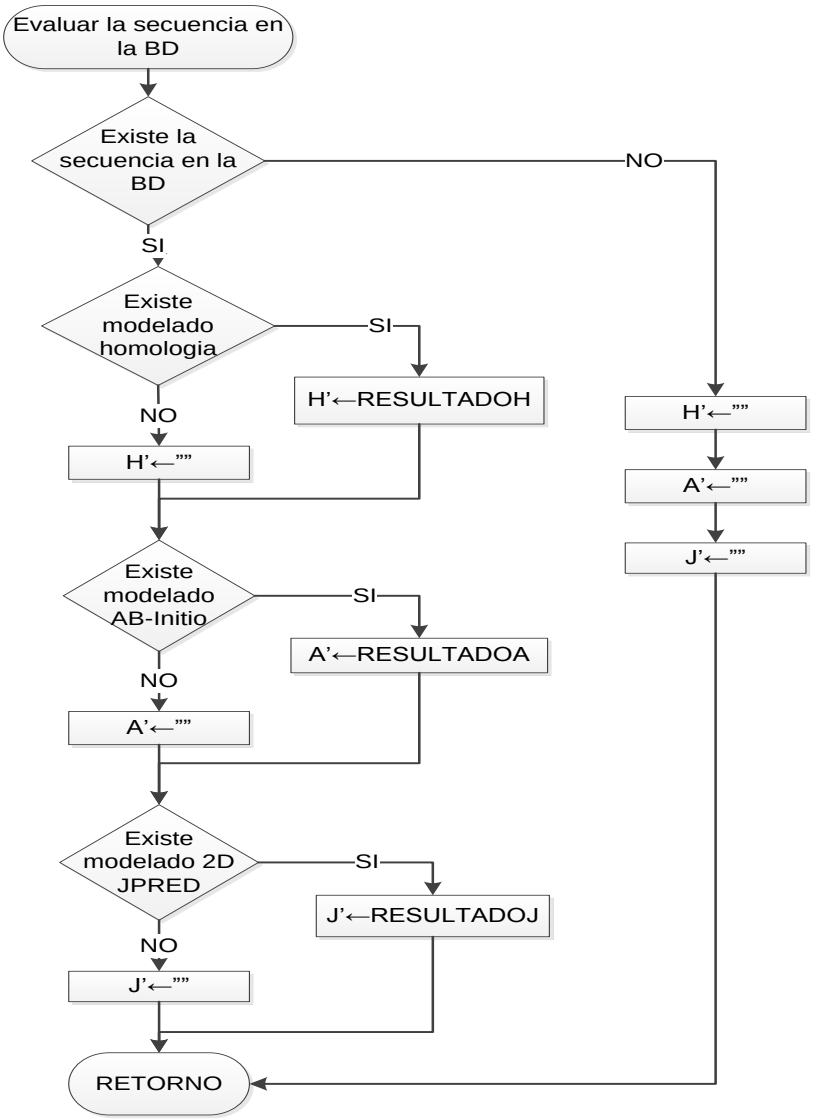


Figura 4. Validación si la proteína a predecir existe en la base de datos del aplicativo Web.

Método de homología.

Para la predicción por el método de homología la interfaz web realiza un llamado al servidor PHP, este se comunica mediante el uso de la función EXEC con un algoritmo realizado con el framework PHANTOMJS. Este algoritmo se encarga de ingresar al servidor PSIBLAST y obtener los resultados del mismo. La función que realiza el servidor web PSIBLAST es usada para buscar posibles secuencias homólogas. Este programa primero realiza un alineamiento entre las diferentes secuencias obtenidas, posteriormente realiza una matriz estándar para calificar los alineamientos realizados. De las secuencias obtenidas en este alineamiento, el programa genera una nueva matriz de sustitución, basándose en las frecuencias de los aminoácidos de las secuencias obtenidas en los alineamientos. Usa esta nueva matriz para realizar otro alineamiento. Esto permite en general encontrar nuevos alineamientos, que son usados para calcular una nueva matriz. El proceso se repite tantas veces como el usuario lo indique, o hasta que ya no se encuentran nuevos alineamientos [14]. Este servidor entrega como resultados un conjunto de secuencias cercanas de acuerdo a los criterios del servidor con la proteína objetivo. También se obtiene la cadena de aminoácidos en formato FASTA de la secuencia más cercana y el alineamiento entre la secuencia objetivo y la cercana. Este último algoritmo retorna los resultados al servidor PHP y este a su vez retorna este resultado a la interfaz WEB para su visualización.

Con los resultados anteriormente obtenidos se procede a hacer uso del servidor CLUSTAL-OMEGA, el cual es un programa de alineamiento de múltiples secuencias que utiliza arboles de guía y técnicas de perfil HMM para generar alineaciones entre diferentes secuencias [15]. Haciendo uso de la misma metodología entradas salidas descrita anteriormente, es decir, petición al servidor, procesamiento de la información, ejecución de algoritmos PHANTOMJS y devolución de resultados, el algoritmo de procesamiento accede al servidor CLUSTAL-OMEGA y únicamente obtiene los resultados arrojados por el mismo. El

resultado arrojado por CLUSTAL-OMEGA, consiste en el alineamiento realizado entre la secuencia objetivo y la secuencia más cercana entregada por el servidor PSIBLAST, a pesar de que este último entrega también una alineación de estructuras su formato no es adecuado para el proceso de predicción por esto es necesario el uso de CLUSTAL-OMEGA.

A partir del alineamiento realizado por el servidor CLUSTAL-OMEGA, se procede a acceder al servidor SWISS-MODEL [16]. Desde la interfaz web se envían los resultados del alineamiento anteriormente obtenido para realizar la predicción de su estructura terciaria. En este servidor se debe tener en cuenta que si la secuencia más cercana obtenida no existe en la base de datos del SWISS-MODEL, la predicción no podrá ser realizada con el alineamiento obtenido por medio del CLUSTAL-OMEGA, en este caso solo se evalúa la predicción de la secuencia objetivo. Posteriormente siendo que se obtengan los resultados del servidor SWISS-MODEL por medio de las secuencias alineadas o la secuencia objetivo, este devolverá como resultado diferentes modelos de la estructura tridimensional obtenida, set de referencia, el estimado de calidad local y el porcentaje de acierto del modelo construido según SWISS-MODEL, es importante destacar que una predicción no se valida por un porcentaje de acierto determinado o un valor especifico sino por los mismos criterios de cada servidor. Con la obtención de este resultado se termina el proceso de predicción por el método de homología.

En la figura 5 se muestra el diagrama de flujo de los pasos a realizar para la obtención del modelo tridimensional a partir de la secuencia de aminoácidos por medio del modelado por homología.

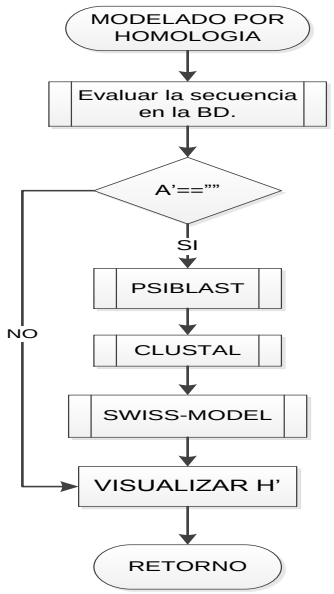


Figura 5. Diagrama de flujo para el modelado por homología.

En la figura 6 se muestran los diagramas de flujos de los pasos realizados por los algoritmos para la obtención del modelo tridimensional por el método de homología.

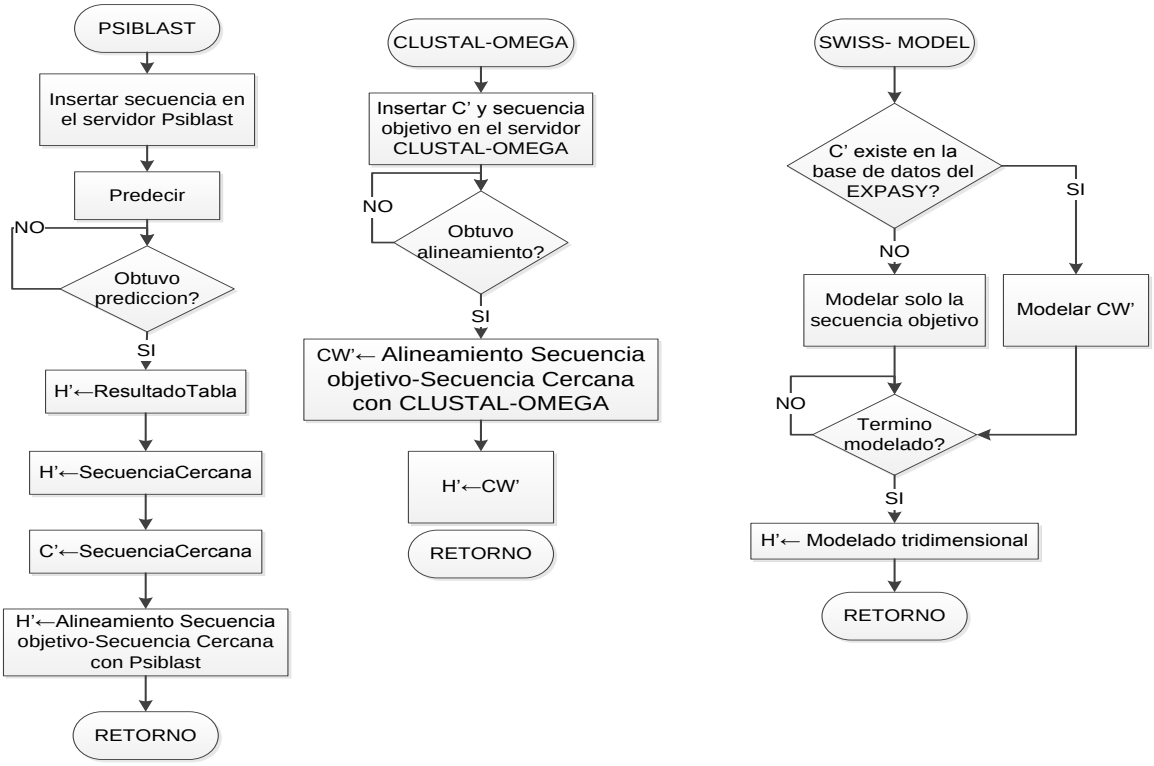


Figura 6. Diagramas de flujo de los pasos realizados para el modelado por homología.

Método AB-INITIO I-TASSER.

Desde la interfaz web se envía información al servidor I-TASSER, en el cual se determina si se encuentra una predicción en cola o este está disponible para una nueva predicción, esto se realizó debido a que este servidor solo deja realizar una predicción al tiempo por dirección IP. Si el servidor está disponible se envía el pedido de la predicción de la proteína objetivo, este responde con el código de trabajo para esta predicción. Este resultado es enviado a la interfaz y este a su vez hace un llamado al servidor PHP para que revise cada cierto intervalo de tiempo, definido en 30 minutos, el correo de resultados dispuesto para la aplicación, esto se debe a que el resultado de I-TASSER es enviado por correo electrónico.

La función que realiza el I-TASSER es generar automáticamente las predicciones de alta calidad de la estructura 3D y la función biológica de las moléculas de proteína a partir de sus secuencias de aminoácidos. El servidor I-TASSER es un banco de trabajo on-line de alta resolución de modelado de la estructura y función de proteínas. Dada una secuencia de la proteína, una salida típica del servidor I-TASSER incluye la predicción de estructura secundaria, prevé la accesibilidad solvente de cada residuo, las proteínas homólogas de plantilla de detectado por roscado y alineaciones de la estructura, modelos estructurales terciarios, estructuras funcionales para la clasificación de las enzimas, los términos de ontología de genes y la proteína ligando sitios de unión. Todas las previsiones están marcadas con una puntuación de confianza que cuenta la precisión de la predicción sin conocer los datos experimentales. Para facilitar las solicitudes especiales de los usuarios finales, el servidor proporciona canales para aceptar especificaciones del usuario acerca de la distancia y el mapa de contacto de los residuos, además también permite al usuario especificar cualquier proteína como plantilla, o excluir a cualquier plantilla de proteínas durante las simulaciones de la estructura de montaje. Para su mayor entendimiento por favor dirigirse a la referencia [17].

Retomando la obtención de los resultados, cada media hora haciendo uso de la función IMAP de PHP y de los algoritmos realizados, se verifica la existencia de los resultados. Si los resultados existen, se retorna la información de los mismos a la interfaz web, si estos no existen se esperará 30 minutos para realizar de nuevo el proceso. Este servidor entrega como resultado varios modelos tridimensionales obtenidos a partir de la secuencia de aminoácidos objetivo con su respectico C-score, donde este C-score es una medida de confianza para la estimación de la calidad de los modelos realizados por el I-TASSER, esta medida se encuentra entre el intervalo de [-5, 2], lo que significa que si el valor es mayor el modelo tiene un alto grado de confianza. También este servidor hace entrega de la estructura secundaria y de una gráfica por modelo donde se puede observar el número de residuos y la distancia estimada que existe entre cada residuo.

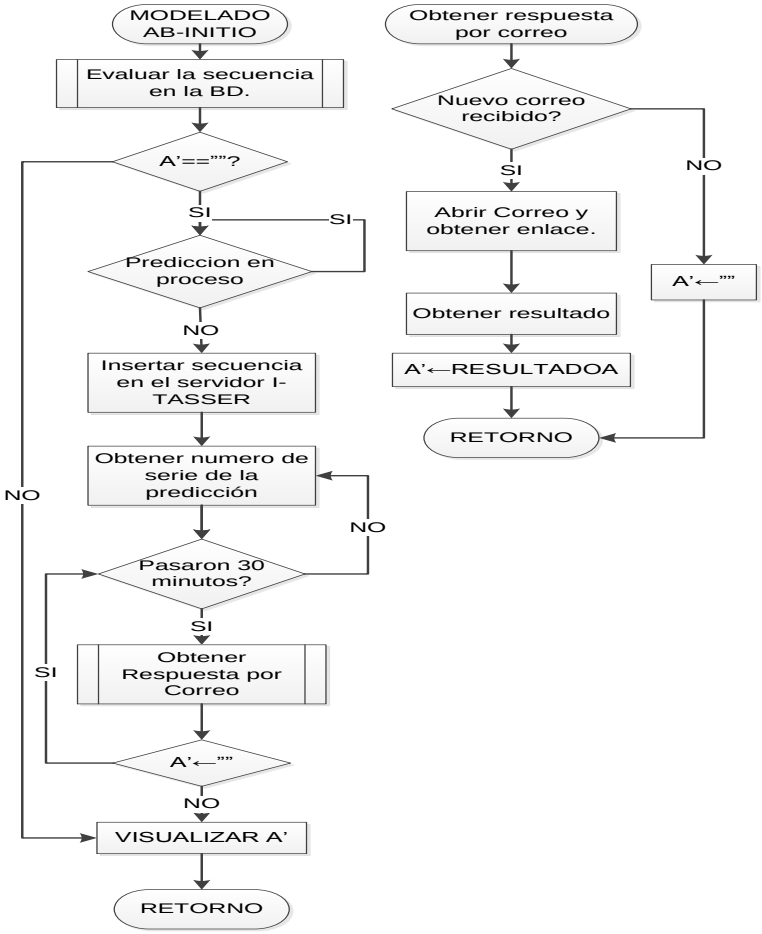


Figura 7. Diagrama de flujo para el modelado tridimensional por AB-Initio.

En la figura 7 se muestra el diagrama de flujo de los pasos a realizar para la obtención del modelo tridimensional a partir de la secuencia de aminoácidos por medio del modelado AB-Initio.

Método predicción 2D JPRED4.

Mediante la interfaz web se envía la secuencia de aminoácidos en formato FASTA al servidor web JPRED4. Posteriormente se procede a validar la finalización de la creación de los resultados para así obtener los datos de la estructura secundaria de la proteína objetivo. Si los datos son obtenidos, estos se envían a la interfaz web para su respectiva visualización. Este modelo entrega como resultado la secuencia objetiva y diferentes anotaciones como son:

- **Lupas_21, Lupas_14, Lupas_28:** son predicciones en espiral de la secuencia. Estos son predicciones binarias para cada ubicación.
- **JNETSOL25, JNETSOL5, JNETSOL0:** Son predicciones solventes de accesibilidad - predicciones binarias de 25%, 5% o 0% de solvente de accesibilidad.
- **JNetPRED** Es la predicción de consenso, donde la predicción de consenso se calcula mediante la predicción de cada método, cada posición y se toma el estado más popular. – las hélices están marcados como tubos rojos y las hojas como flechas de color verde oscuro, este es el resultado que se toma como resultado final del servidor.
- **JNetCONF** Es la estimación de la confianza para la predicción. Los valores altos significan una alta confianza.
- **JNetHMM** predicción basada en los modelos ocultos de Markov (HMM). Este modelo se visualiza como una máquina de estados finitos, donde genera una secuencia de proteína de aminoácidos a medida que avanza a través de una serie de estados. Cada estado tiene una mesa de probabilidades de emisión de aminoácidos similares a los descritos en un modelo de perfil. También hay probabilidades de transición para pasar de

un estado a otro. [18] - hélices están marcados como tubos rojos y hojas como flechas de color verde oscuro.

- **JNETPSSM** predicción basada PSSM. Un PSSM, o matriz de posición específica de puntuación, es un tipo de matriz de puntuación utilizado en búsquedas en la que las puntuaciones de sustitución de aminoácidos se dan por separado para cada posición en una alineación de proteínas en múltiples secuencias. [19] - hélices están marcados como tubos rojos y hojas como flechas de color verde oscuro.
- **JNETJURY A '*'** en esta anotación indica que el JNETJURY se invocó para racionalizar significativamente diferentes predicciones primarias.

En la figura 8 se observa el diagrama de flujo de los pasos a realizar para la obtención del modelo secundario a partir de la secuencia de aminoácidos por medio del modelado JPRED4.

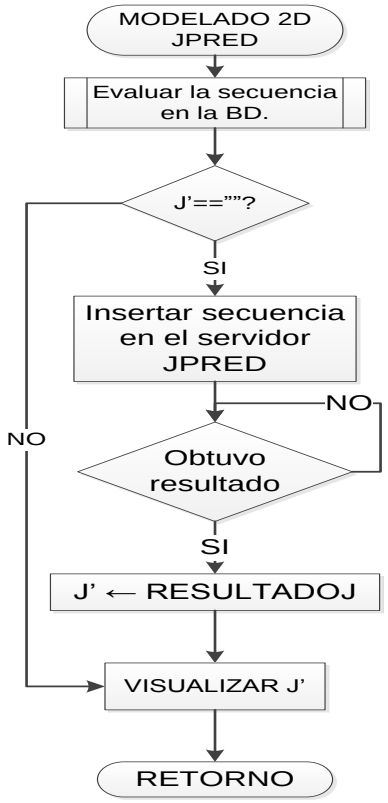


Figura 8. Diagrama de flujo para el modelado secundario por JPRED4.

5. ANÁLISIS Y RESULTADOS.

En este capítulo se describirán y analizaran los resultados obtenidos una vez seguida la metodología del capítulo anterior.

Una vez almacenado y configurado el aplicativo web sobre un servidor local Windows, en este caso un computador marca HP con 4GB de memoria RAM, 500GB de disco duro, procesador Intel Core i7 de 2.13GHz y con sistema operativo Windows 7 Home Premium Service Pack 1 de 64 bit se procede a inicializar el servidor. Para su correcto funcionamiento se inicia el aplicativo XAMPP, seleccionando los módulos de APACHE y MySQL.

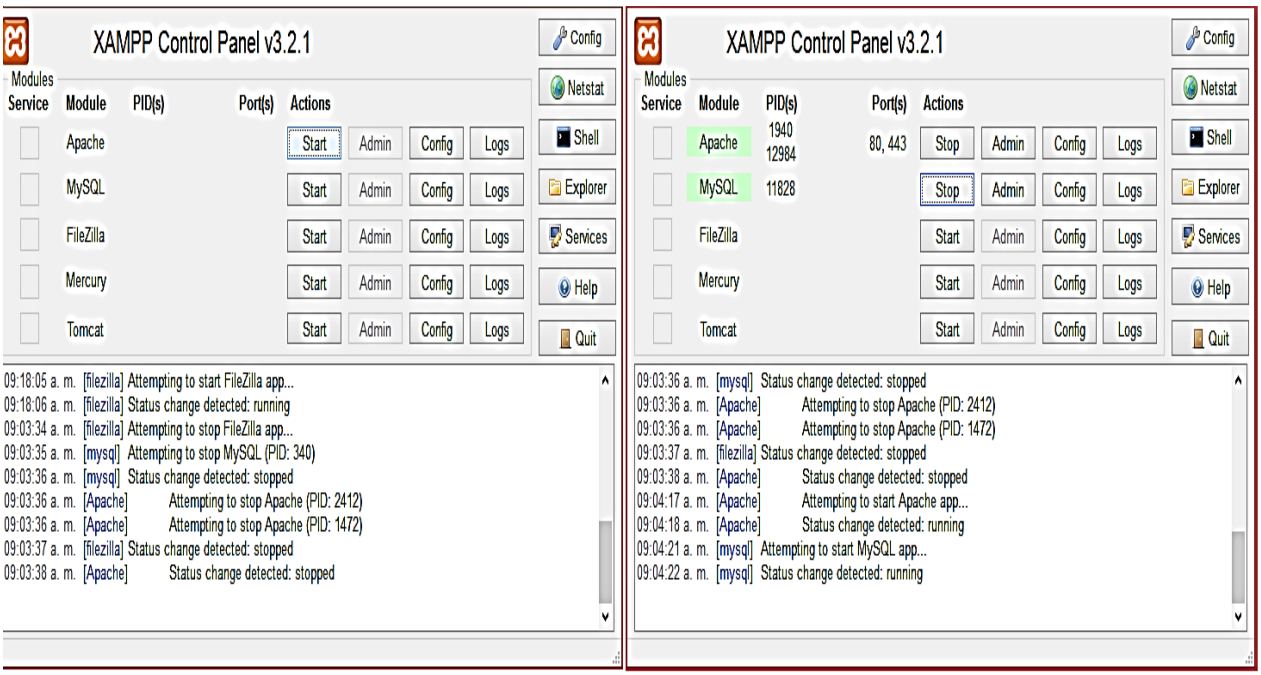


Figura 9. Ejecución XAMPP.

Al dar inicio a los módulos antes mencionados del XAMPP, se procede a dirigirse a un navegador Web para la ejecución del aplicativo web. Para su ejecución se abre la ubicación donde están almacenados todos los algoritmos realizados para

la predicción, esta ubicación puede estar determinada por una url local o un dominio. Como en este caso el aplicativo se encuentra almacenado en un computador portátil, el cual sirve como servidor local la dirección de URL es **http://localhost/predictorestructuras/web/**

Figura 10. Inicio del Aplicativo WEB

Al ingresar al aplicativo se procede a insertar la secuencia de aminoácidos en formato FASTA. También se selecciona que tipo de modelo para la predicción de proteínas se desea realizar (Homología, Ab-Initio, Estructura Secundaria). Después de realizado lo anterior se procede a hacer click sobre el botón **Predecir**, de esta forma se inicia la ejecución de los diferentes algoritmos necesarios para la predicción seleccionada.

5.1. Pruebas.

5.1.1. Primera Prueba

Para la primera prueba se utilizó la secuencia de aminoácidos de la proteína **Polifenoloxidasa del lulo** la cual se encuentra en su formato FASTA. Esta secuencia fue obtenida con ayuda del centro de biología molecular (Cenbiotep), el cual pertenece a la Universidad Tecnológica de Pereira. Se utiliza esta secuencia

de proteína debido al interés del centro de biología molecular en la misma y la poca información que existe sobre dicha proteína, esto es debido a que el lulo solo es importante en la zona andina y no se le ha dado mucha relevancia al mejoramiento de defectos en la fruta por medio de la modificación genética. Con esta aplicación se busca ayudar en la obtención de información para eliminar el pardeamiento de la fruta, además de brindar una herramienta que pueda predecir cualquier estructura de proteína.

LGSTSKPSQLFHHGKRNKTFKVSCKVTNNNGDQNQNGVDRRNVLLGLGGMYGVANAIPSASSATPAPPPDLSA
CRKANLLVVCLICFANPVSVCILGLASVCTHPSVNSTPILGSI-
PISINETWRERVHTMLLCASLTASMEGWNKIPLFTKFPSS-
LTKLRIRQPAHAADEECIAKYNLAISGMKDLDKTEPSNPYWAFKQQANIHCAYCNGAYAIG-QSVTSH-
LMAFLPCSIHRWYLSTRESLGYKSSTDPTFALPYWNWDHPKGMRMPPMFDREGTALYDESERNPNR-
SVNGNRYMDLGSIRRPSPNNLEIQLMSNN-L-VPSNGN-
CTVNGTMLHVLGCLQVARLMFSGITLEAPGTIERPSHTVSHTHLGLVQSRGYNS-P-W-
NGHTVRDMGHFYSAGLGPRIFFCHHGNVDRMWSEWKRESGGKRRDLHKDWLNSEFFFYD-
RQKPLPCESPRLFGHQEDGVSITHQCQHRGVTSSQKQRPQLKQEAKATAGKVNASSLPPASKVFPLAKLDKAISF
SINRPASSRSQQEKNEQEEMLTFSDIKYDNREYIRFDVFLNVDKNVNADELDKAEFAGSYTSLPHVHRAGDNNH
VATATLRLAVTELLEDIALEYENTIAVTLVPKKDGEGISIGGVEITLADC-L

Figura 11. Secuencia de aminoácidos de la polifenoloxidasa del lulo en formato FASTA.

Obtenida la secuencia esta se inserta en el campo donde la aplicación la requiere y se procede a seleccionar que tipo de predicción se desea realizar. En este caso se seleccionaron las tres opciones (HOMOLOGIA 3D, AB INITIO 3D, JPRED 2D) para obtener los resultados de esta secuencia en los tres modelos diferentes. Se procedió a dar click sobre el botón predecir para la obtención de los resultados. El tiempo para la obtención de estos resultados puede llegar a ser muy variable, esto depende en muchos casos del tamaño de la secuencia de aminoácidos insertada, también se debe tener en cuenta que los servidores a los que se tiene acceso pueden llegar a tener colas de espera, lo cual hace que el tiempo de respuesta sea mucho mayor. Por lo tanto no se tienen tiempos de respuestas determinados.

Modelado por Homología.

Después de la espera de alrededor de 5 minutos se obtuvieron los resultados de la secuencia de aminoácidos por medio de este modelado. Los resultados obtenidos por medio de este modelado se dividen en tres partes importantes, como lo son: **PSIBLAST, CLUSTAL-OMEGA** y **SWISS-MODEL**.

El servidor web **PSIBLAST,** arrojó como resultados la secuencia más cercana obtenida con respecto a la secuencia insertada.

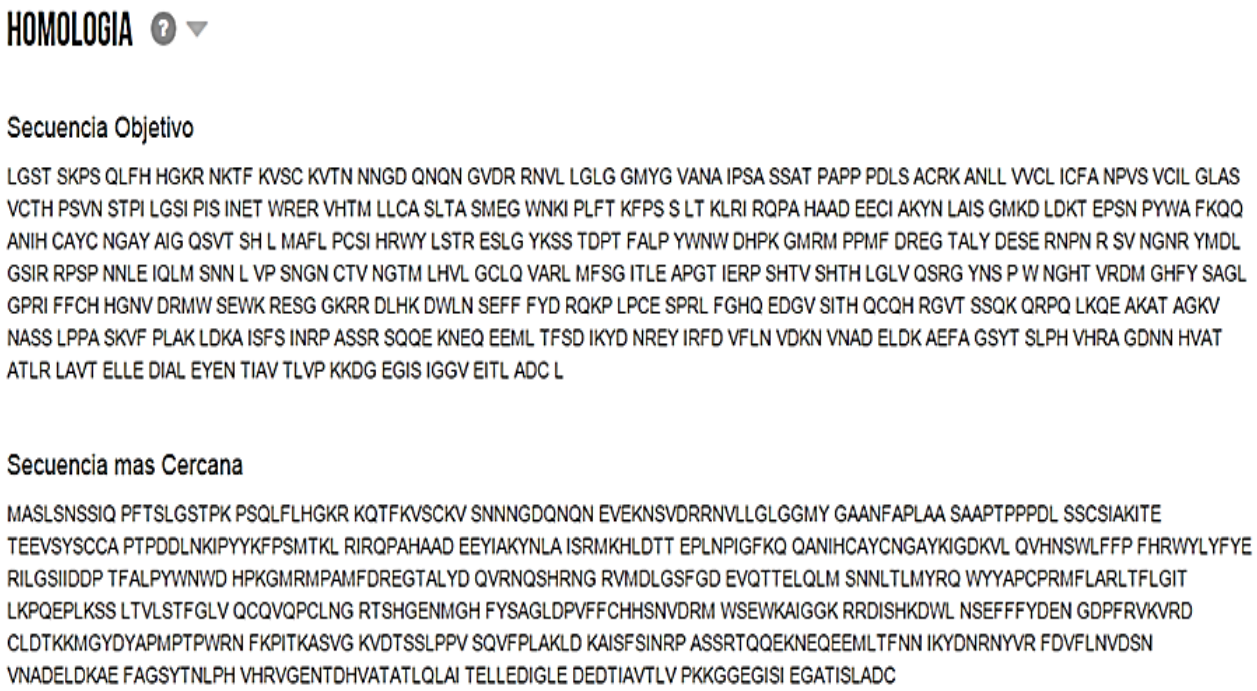
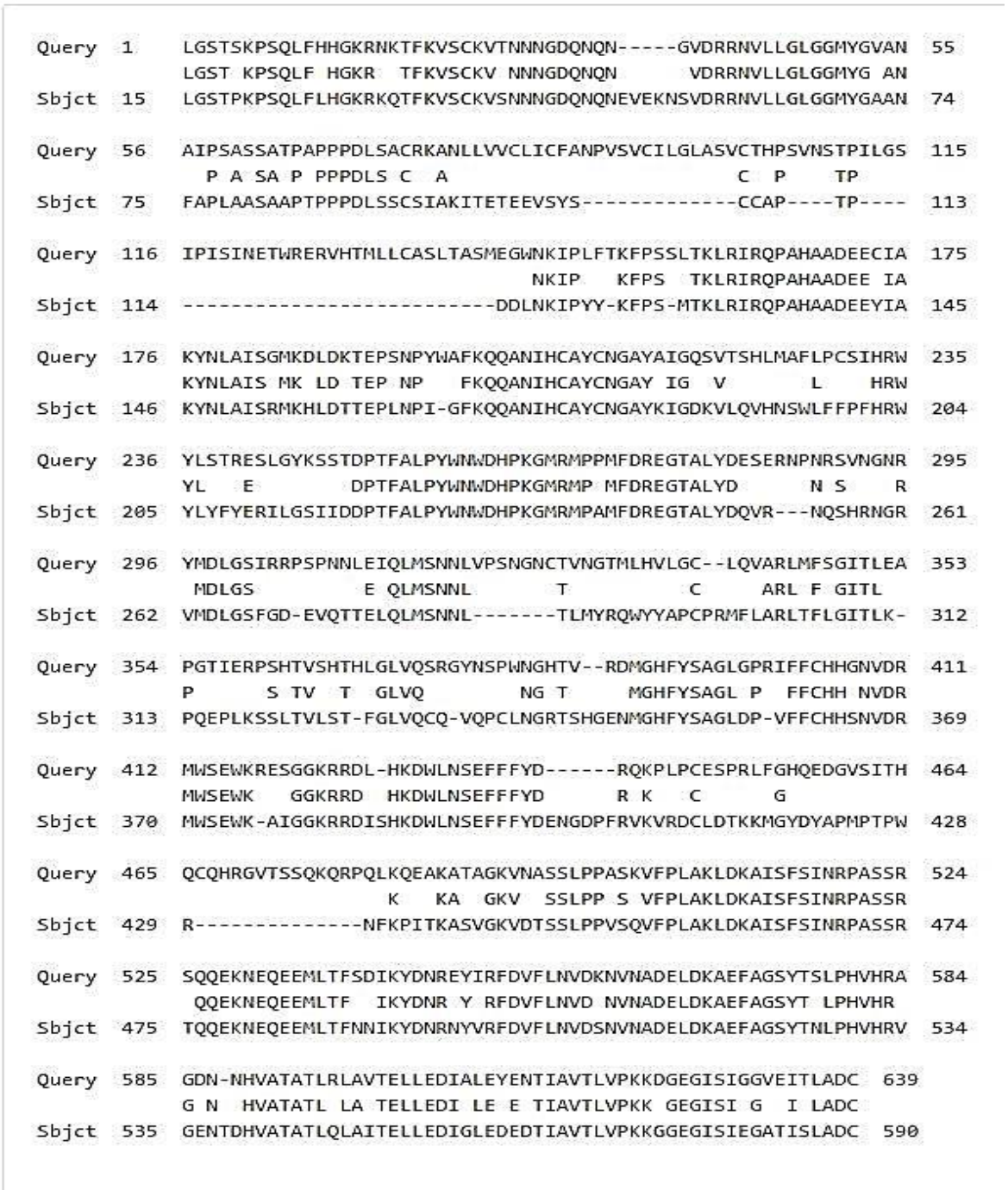


Figura 12. Secuencia objetivo, secuencia más cercana a la secuencia objetivo.

Otros resultados de suma importancia para el entendimiento de la proteína arrojados por el PSIBLAST son el alineamiento entre la secuencia insertada y la secuencia más cercana. Este alineamiento es de gran importancia, debido a que se puede visualizar que zonas son idénticas o tienen similitud. Por ultimo este servidor da como resultado una tabla de secuencias de proteínas más cercanas.

En esta tabla se muestra el nombre de las secuencias, el puntaje máximo de similitud con la secuencia objetivo, entre otros datos importantes.

Alineamiento con secuencia mas cercana



```
Query  1    LGSTSKPSQLFHHGKRNKTFKVSCKVTNNNGDQNQN-----GVDRRNVLLGLGGMYGVAN  55
            LGST KPSQLF HGKR  TFKVSCKV NNNGDQNQN      VDRRNVLLGLGGMYG AN
Sbjct  15   LGSTPKPSQLFLHGKRKQTFKVSCKVSNNNGDQNQNEVEKNSVDRRNVLLGLGGMYGAAN  74

Query  56   AIPSASSATPAPPPDLSACRKANLLVVCLICFANPVSVCILGLASVCTHPSVNSTPILGS  115
              P A SA P PPPDLS C  A                          C  P    TP
Sbjct  75   FAPLAASAAPTPPPDLSSCSIAKITETEEVSYS-------------CCAP----TP----  113

Query  116  IPISINETWRERVHTMLLCASLTASMEGWNKIPLFTKFPSSLTKLRIRQPAHAADEECIA  175
                                       NKIP   KFPS  TKLRIRQPAHAADEE IA
Sbjct  114  --------------------------DDLNKIPYY-KFPS-MTKLRIRQPAHAADEEYIA  145

Query  176  KYNLAISGMKDLDKTEPSNPYWAFKQQANIHCAYCNGAYAIGQSVTSHLMAFLPCSIHRW  235
            KYNLAIS MK LD TEP NP   FKQQANIHCAYCNGAY IG  V       L    HRW
Sbjct  146  KYNLAISRMKHLDTTEPLNPI-GFKQQANIHCAYCNGAYKIGDKVLQVHNSWLFFPFHRW  204

Query  236  YLSTRESLGYKSSTDPTFALPYWNWDHPKGMRMPPMFDREGTALYDESERNPNRSVNGNR  295
            YL   E        DPTFALPYWNWDHPKGMRMP MFDREGTALYD      N S    R
Sbjct  205  YLYFYERILGSIIDDPTFALPYWNWDHPKGMRMPAMFDREGTALYDQVR---NQSHRNGR  261

Query  296  YMDLGSIRRPSPNNLEIQLMSNNLVPSNGNCTVNGTMLHVLGC--LQVARLMFSGITLEA  353
             MDLGS         E QLMSNNL       T          C     ARL F GITL
Sbjct  262  VMDLGSFGD-EVQTTELQLMSNNL-------TLMYRQWYYAPCPRMFLARLTFLGITLK-  312

Query  354  PGTIERPSHTVSHTHLGLVQSRGYNSPWNGHTV--RDMGHFYSAGLGPRIFFCHHGNVDR  411
            P      S TV  T  GLVQ        NG T     MGHFYSAGL P  FFCHH NVDR
Sbjct  313  PQEPLKSSLTVLST-FGLVQCQ-VQPCLNGRTSHGENMGHFYSAGLDP-VFFCHHSNVDR  369

Query  412  MWSEWKRESGGKRRDL-HKDWLNSEFFFYD------RQKPLPCESPRLFGHQEDGVSITH  464
            MWSEWK   GGKRRD  HKDWLNSEFFFYD      R K   C      G
Sbjct  370  MWSEWK-AIGGKRRDISHKDWLNSEFFFYDENGDPFRVKVRDCLDTKKMGYDYAPMPTPW  428

Query  465  QCQHRGVTSSQKQRPQLKQEAKATAGKVNASSLPPASKVFPLAKLDKAISFSINRPASSR  524
                          K   KA  GKV  SSLPP S VFPLAKLDKAISFSINRPASSR
Sbjct  429  R--------------NFKPITKASVGKVDTSSLPPVSQVFPLAKLDKAISFSINRPASSR  474

Query  525  SQQEKNEQEEMLTFSDIKYDNREYIRFDVFLNVDKNVNADELDKAEFAGSYTSLPHVHRA  584
             QQEKNEQEEMLTF  IKYDNR Y RFDVFLNVD NVNADELDKAEFAGSYT LPHVHR
Sbjct  475  TQQEKNEQEEMLTFNNIKYDNRNYVRFDVFLNVDSNVNADELDKAEFAGSYTNLPHVHRV  534

Query  585  GDN-NHVATATLRLAVTELLEDIALEYENTIAVTLVPKKDGEGISIGGVEITLADC  639
            G N  HVATATL LA TELLEDI LE E TIAVTLVPKK GEGISI G  I LADC
Sbjct  535  GENTDHVATATLQLAITELLEDIGLEDEDTIAVTLVPKKGGEGISIEGATISLADC  590
```

Figura 13. Alineamiento entre secuencia objetivo-secuencia más cercana.

Tabla de proteinas cercanas

Descripcion	Puntaje Maximo	Puntaje Total	Query Cover	Valor E	Identidad	Accession
polyphenol oxidase [Solanum melongena]	617	617	99%	0.0	58%	ACR61398.1
chloroplast polyphenol oxidase [Solanum melongena]	616	616	99%	0.0	58%	AFJ79640.1
chloroplast polyphenol oxidase [Solanum melongena]	615	615	99%	0.0	57%	AFJ79641.1
chloroplast polyphenol oxidase [Solanum melongena]	613	613	99%	0.0	57%	AFJ79643.1
chloroplast polyphenol oxidase precursor [Solanum melongena]	613	613	99%	0.0	56%	ADG56700.1
chloroplast polyphenol oxidase [Solanum melongena]	613	613	99%	0.0	57%	AFJ79639.1
polyphenol oxidase [Solanum tuberosum]	607	607	99%	0.0	57%	AAA85122.1
PREDICTED: polyphenol oxidase B, chloroplastic [Solanum lycopersicum]	605	605	99%	0.0	56%	XP_004246032.1
chloroplast polyphenol oxidase precursor [Solanum melongena]	592	592	99%	0.0	55%	ACT22523.1
PREDICTED: polyphenol oxidase B, chloroplastic-like [Solanum tuberosum]	590	590	99%	0.0	57%	XP_006365382.1

Figura 14. Tabla de proteínas más cercanas.

En la figura 14 se observan las 10 secuencias más cercanas a la secuencia objetivo. Si se desea aumentar la visualización de secuencias en la tabla, se debe proceder a hacer click sobre el icono que se encuentra en la parte superior derecha de la tabla, el cual tiene una forma redonda y un signo “+”. Al hacer click sobre este icono se obtendrán las 100 secuencias más cercanas. También se puede observar en esta figura el valor de identidad de cada secuencia de aminoácidos con respecto a la secuencia objetivo, con esta secuencia este valor es un poco bajo, esto se debe a la información inexistente acerca del lulo. En este caso la secuencia más cercana obtenida con más alto valor de puntaje es de la polifenoloxidasa de la Solanum melongena o más conocida como la berenjena. Es importante destacar que el valor de identidad solo define el porcentaje de aminoácidos que contiene la secuencia homologa con respecto a la secuencia objetivo. Para determinar la mejor predicción lo que realmente importa es el puntaje máximo siendo este quien define la mejor secuencia a utilizar, pues la homología no solo se da por la similitud de caracteres sino también por otros factores como la distancia entre residuos.

Continuando con el proceso, el servidor CLUSTAL OMEGA entrega como resultado el alineamiento para su posterior utilización en el servidor SWISS-MODEL. En la figura 15 se muestra el alineamiento. Para este caso hay cuatro tipos de símbolos que puede arrojar el alineamiento:

- * - indica las posiciones que contienen un único residuo totalmente conservado.
- : - indica que los residuos comparten propiedades fuertemente similares.
- . – indica que los residuos compartes propiedades débilmente similares.
- - Indican que los residuos no comparten ninguna propiedad.

Alineamiento con clustal

```
CLUSTAL O(1.2.1) multiple sequence alignment


EMBOSS_001      --------------LGSTSKPSQLFHHGKRNKTFKVSCKVTNNNGDQNQN-----GVDRR
EMBOSS_002      MASLSNSSIQPFTSLGSTPKPSQLFLHGKRKQTFKVSCKVSNNNGDQNQNEVEKNSVDRR
                          **** ****** ****::********:*********     .****

EMBOSS_001      NVLLGLGGMYGVANAIPSASSATPAPPPDLSACRKANLLVVCLICFANPVSVCILGLASV
EMBOSS_002      NVLLGLGGMYGAANFAPLAASAAPTPPPDLSSCSIAKITETEEVSY-------------S
                **********.**  * *:**:*:******:*  *::  .  :.:

EMBOSS_001      CTHPSVNSTPILGSI-PISINETWRERVHTMLLCASLTASMEGWNKIPLFTKFPSS-LTK
EMBOSS_002      C--------------------------------CA---PTPDDLNKIPYY-KF--PSMTK
                *              *                  **    : :  **** : **    :**

EMBOSS_001      LRIRQPAHAADEECIAKYNLAISGMKDLDKTEPSNPYWAFKQQANIHCAYCNGAYAIG-Q
EMBOSS_002      LRIRQPAHAADEEYIAKYNLAISRMKHLDTTEPLNPI-GFKQQANIHCAYCNGAYKIGDK
                ************ ******** **.**.*** **  .*************** ** :

EMBOSS_001      SVTSH-LMAFLPCSIHRWYLSTRESLGYKSSTDPTFALPYWNWDHPKGMRMPPMFDREGT
EMBOSS_002      VLQVHNSWLFF--PFHRWYLYFYERILGSIIDDPTFALPYWNWDHPKGMRMPAMFDREGT
                 :  *    *:   :*****   * :  .   ****************** *******

EMBOSS_001      ALYDESERNPNR-SVNGNRYMDLGSIRRPSPNNLEIQLMSNN-L-VPSNGN-CTVNGTML
EMBOSS_002      ALYDQVRNQSHR----NGRVMDLGSFGDEV-QTTELQLMSNNLTLMYRQWY--------Y
                ****: ..: .**     * *****:     :. *:******   :  :  *

EMBOSS_001      HVLGCLQVARLMFSGITLEAPGTIERPSHTVSHTHLGLVQSRGYNS-P-W-NGHTVRDMG
EMBOSS_002      APCPRMFLARLTFLGITLKPQEPLKSSLTVL--STFGLVQCQVQPCL--NGRTSHGENMG
                     : :*** * ****:    ::    .:  : :****.:   .  *  .    .:**

EMBOSS_001      HFYSAGLGPRIFFCHHGNVDRMWSEWKRESGGKRRDLHKDWLNSEFFFYD-RQKPLPCES
EMBOSS_002      HFYSAGLDP-VFFCHHSNVDRMWSEWKAIGGKRRDISHKDWLNSEFFFYDENGDPFRVKV
                ******* * :*****.*********  .* :*   ************ . .*:  :

EMBOSS_001      PRLFGHQEDGVSITHQCQHRGVTSSQKQRPQLKQEAKATAGKVNASSLPPASKVFPLAKL
EMBOSS_002      RDCLDTKKMGYDYA---------PMPTPWRNFKPITKASVGKVDTSSLPPVSQVFPLAKL
                   :  :: * . :            .   ::*  :**:.***::*****.*:*******

EMBOSS_001      DKAISFSINRPASSRSQQEKNEQEEMLTFSDIKYDNREYIRFDVFLNVDKNVNADELDKA
EMBOSS_002      DKAISFSINRPASSRTQQEKNEQEEMLTFNNIKYDNRNYVRFDVFLNVDSNVNADELDKA
                ***************:************.:******:*:*********.**********

EMBOSS_001      EFAGSYTSLPHVHRAGDN-NHVATATLRLAVTELLEDIALEYENTIAVTLVPKKDGEGIS
EMBOSS_002      EFAGSYTNLPHVHRVGENTDHVATATLQLAITELLEDIGLEDEDTIAVTLVPKKGGEGIS
                *******.******.*:* :*******:**:*******.** *:********** *****

EMBOSS_001      IGGVEITLADC-L
EMBOSS_002      IEGATISLADC--
                * *. *:*****
```

Figura 15. Alineamiento con CLUSTAL-OMEGA

Para terminar con el modelado por homología, se obtuvo la respuesta del servidor SWISS-MODEL, el cual entrega como resultados el **estimado de calidad local** que es una propiedad de alineación de la secuencia objetivo, lo que esto indica es que a cada residuo se le asigna una puntuación de fiabilidad entre 0 y 1 el cual describe la similitud esperada para la secuencia objetivo [20]**, set de referencia** el cual es una estimación de calidad de un modelo comparándolo con secuencias que tengan conocidas su estructura por medio de la cristalografía de rayos x [21]**, modelo construido** y **el porcentaje de acierto del modelo construido.** Para este alineamiento entre las secuencias de aminoácidos este servidor entregó 3 modelos construidos.

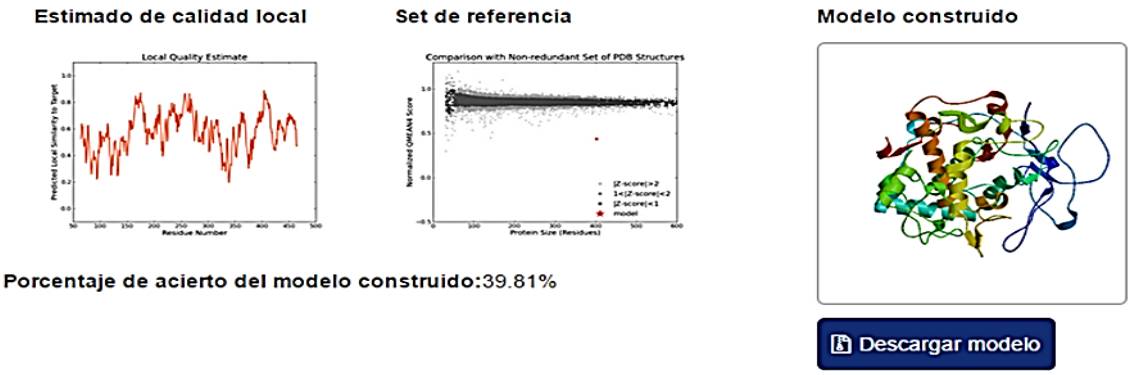


Figura 16. Modelo tridimensional N°1 de la secuencia de aminoácidos de la polifenoloxidasa del lulo.

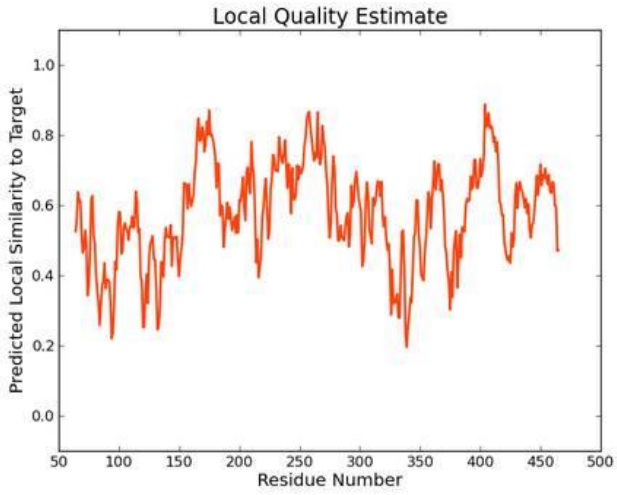


Figura 17. Modelo tridimensional N°1 estimado de calidad local de la secuencia de aminoácidos de la polifenoloxidasa del lulo.

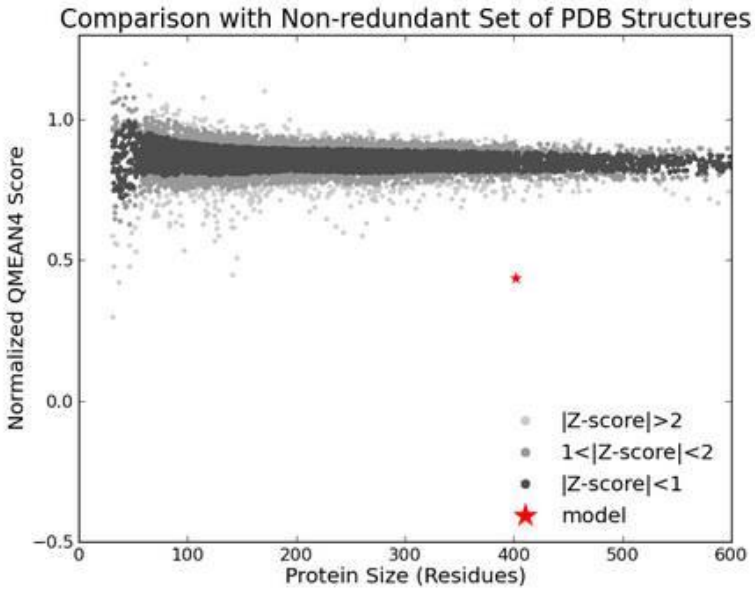


Figura 18. Modelo tridimensional N°1 set de referencia de la secuencia de aminoácidos de la polifenoloxidasa del lulo.

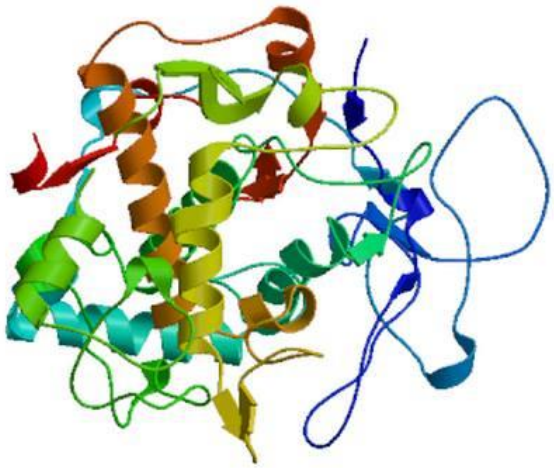


Figura 19. Modelo tridimensional N°1 modelo construido de la secuencia de aminoácidos de la polifenoloxidasa del lulo.

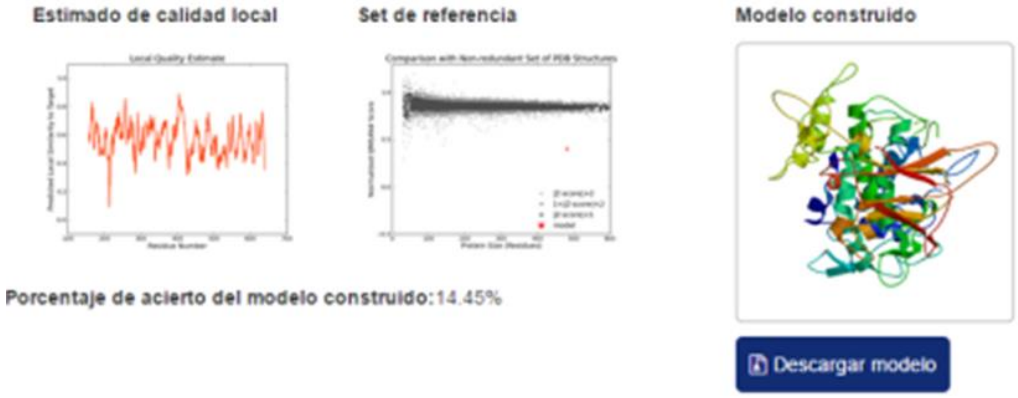


Figura 20. Modelo tridimensional N°2 de la secuencia de aminoácidos de la polifenoloxidasa del lulo.

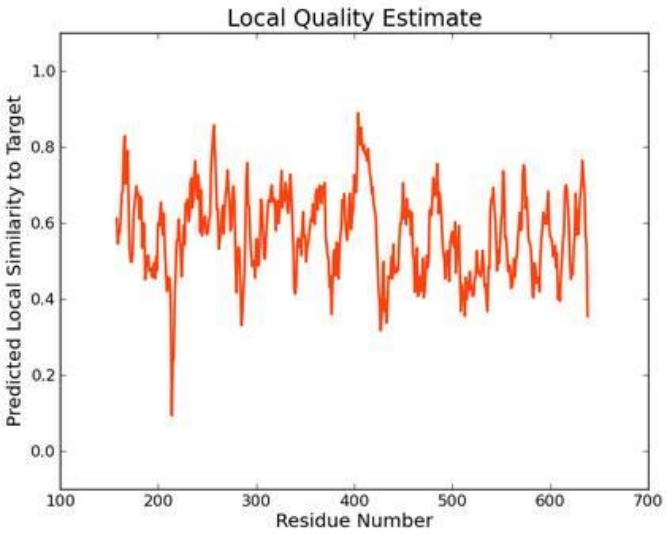


Figura 21. Modelo tridimensional N°2 estimado de calidad local de la secuencia de aminoácidos de la polifenoloxidasa del lulo.

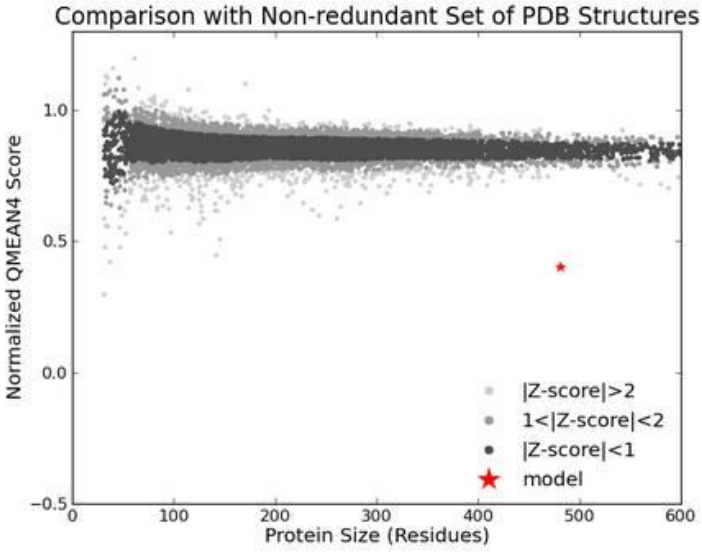


Figura 22. Modelo tridimensional N°2 set de referencia de la secuencia de aminoácidos de la polifenoloxidasa del lulo.

Figura 23. Modelo tridimensional N°2 modelo construido de la secuencia de aminoácidos de la polifenoloxidasa del lulo.

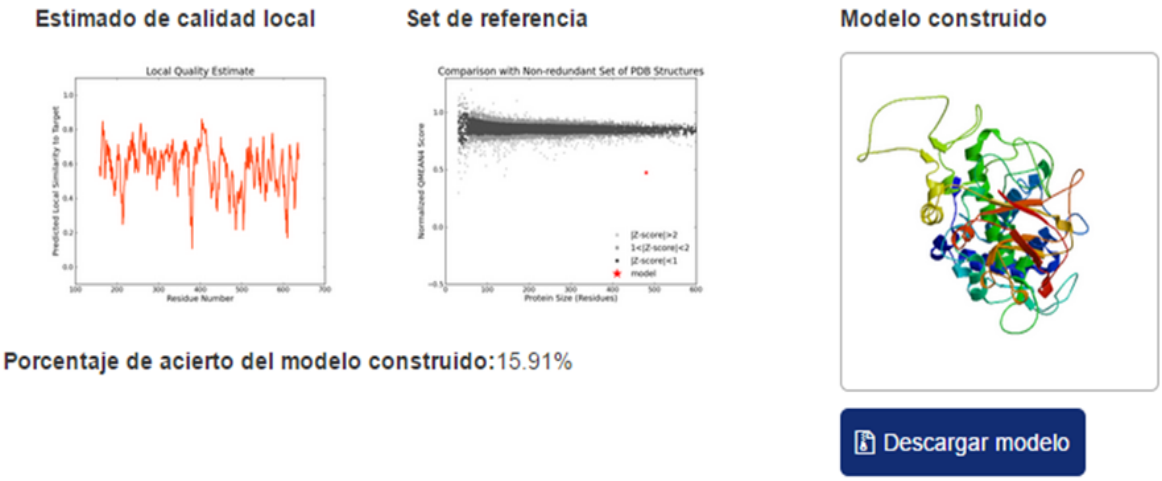


Figura 24. Modelo tridimensional N°3 de la secuencia de aminoácidos de la polifenoloxidasa del lulo.

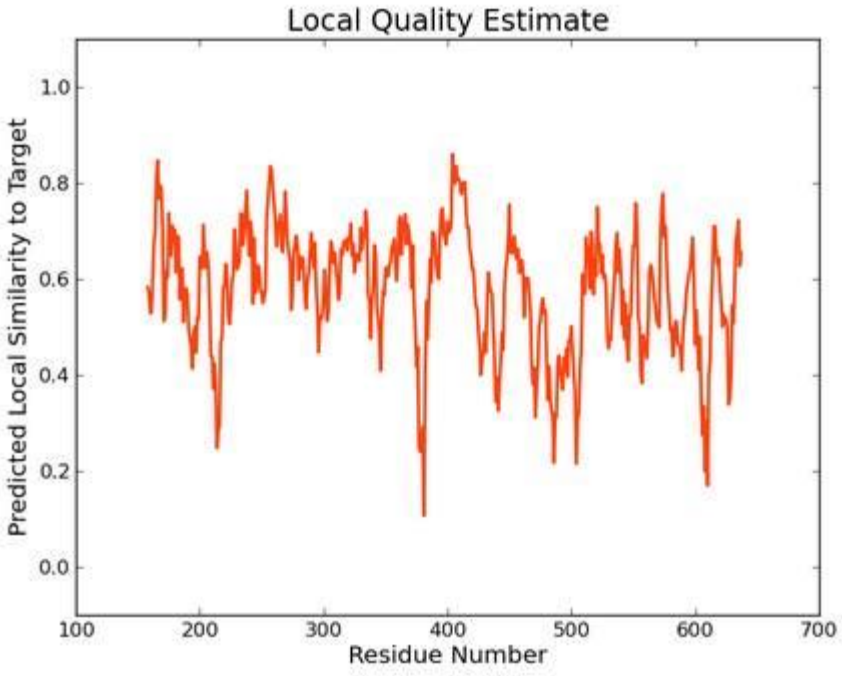


Figura 25. Modelo tridimensional N°3 estimado de calidad local de la secuencia de aminoácidos de la polifenoloxidasa del lulo.

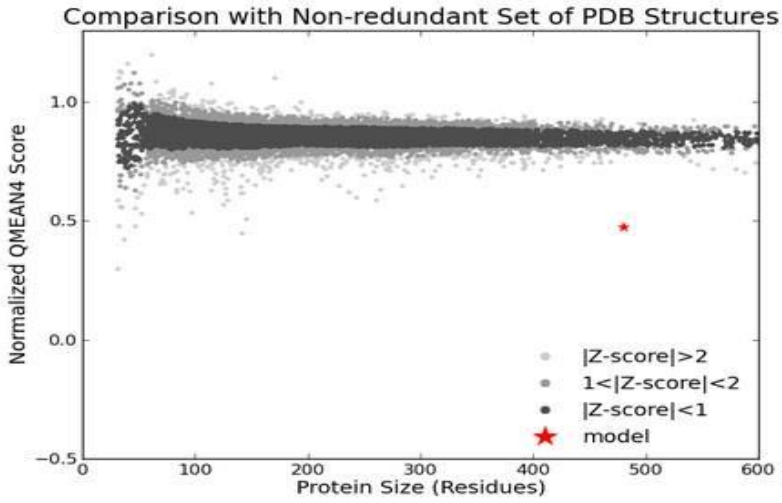


Figura 26. Modelo tridimensional N°3 set de referencia de la secuencia de aminoácidos de la polifenoloxidasa del lulo.

Figura 27. Modelo tridimensional N°3 modelo construido de la secuencia de aminoácidos de la polifenoloxidasa del lulo.

Como se puede observar en los tres modelos tridimensionales construidos a partir de su alineamiento, se puede concluir que los modelos obtenidos tienen un bajo porcentaje de acierto, el máximo porcentaje de acierto entregado por este servidor es del 39.81%, esto se debe a como la secuencia objetivo es decir la secuencia de la proteína polifenoloxidasa del lulo no tiene una secuencia con altos índices de homología además de que la información de esta es casi nula, por lo tanto el servidor ejecuta el algoritmo de modelado con la secuencia con más alta homogeneidad posible aunque esta cuente con una secuencia significantemente diferente. Por esta razón el porcentaje de acierto es bajo.

La aplicación de predicción de proteínas cuenta con un botón llamado descargar modelo, este botón lo que permite es ejecutar la descarga del modelo construido. Este modelo se descarga en formato .pdb. Esto se realizó con el fin de la utilización de este modelo tridimensional en aplicativos como JMOL para su posterior estudio. El aplicativo permite una mayor interacción, desplazamientos en

3D, entre otros diferentes intereses para el área de la Bioinformática tales como observar los ligamentos de hidrogeno.

Figura 28. Modelo tridimensional N°1 en JMOL de la secuencia de aminoácidos de la polifenoloxidasa del lulo.

Modelado AB-Initio.

La obtención de la respuesta del servidor que realiza este modelado, se puede llegar a demorar dependiendo de la cola de trabajos pendientes que existan en este servidor. Los tiempos de respuesta son demasiados variables.

Este tipo de modelado entrega como resultado los modelos más cercanos con su respectivo C-score. También este servidor hace entrega de la estructura secundaria y la gráfica donde se puede observar el número de residuos y la distancia estimada. Con la estructura secundaria se puede visualizar cuando un residuo presenta la característica de ser una Hélice (H), Hebra(S) o Espiral(C). En cuanto a la gráfica obtenida se visualiza, que el valor medio obtenido es muy alto, por esta razón el modelo tridimensional construido a partir de la secuencia de aminoácidos de la polifenoloxidasa del lulo tiende a dar valores de confianza o C-score muy altos.

AB INITIO

Estructura secundaria

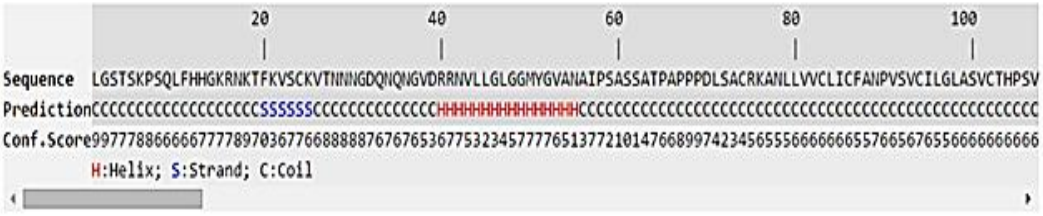
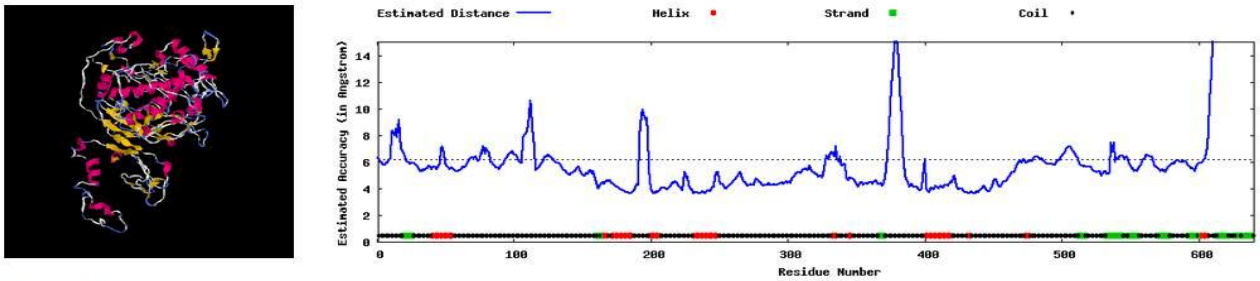


Figura 29. Estructura secundaria obtenido con I-TASSER de la secuencia de aminoácidos de la polifenoloxidasa del lulo.

MODELO 1

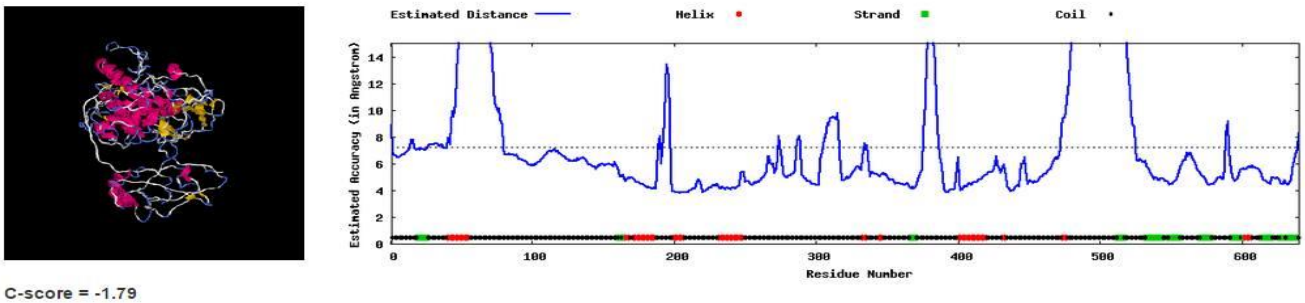


C-score=-1.24

Descargar modelo

Figura 30. Modelo tridimensional I-Tasser N°1 de la secuencia de aminoácidos de la polifenoloxidasa del lulo.

MODELO 2

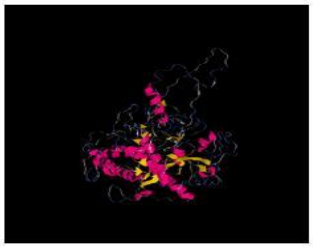


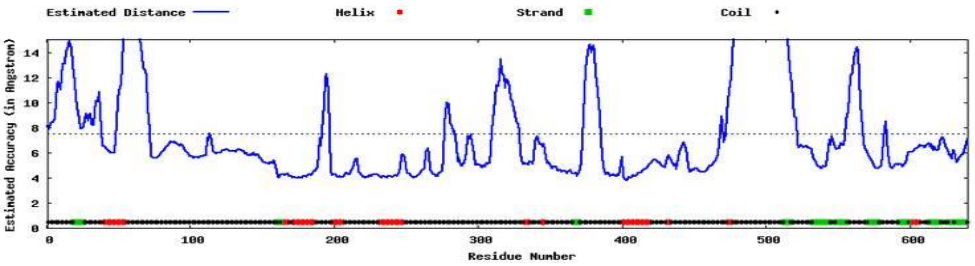
C-score = -1.79

Descargar modelo

Figura 31. Modelo tridimensional I-Tasser N°2 de la secuencia de aminoácidos de la polifenoloxidasa del lulo.

MODELO 3



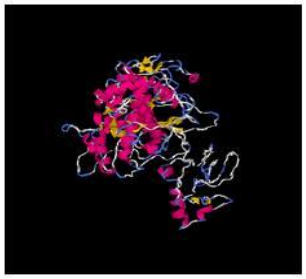


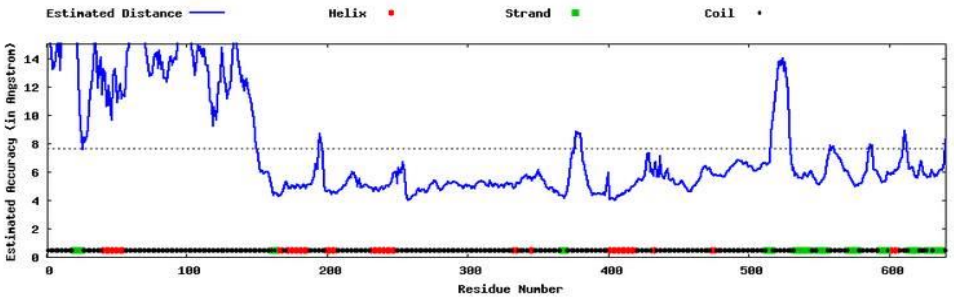
C-score = -1.86

Descargar modelo

Figura 32. Modelo tridimensional I-Tasser N°3 de la secuencia de aminoácidos de la polifenoloxidasa del lulo.

MODELO 4



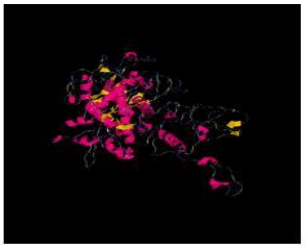


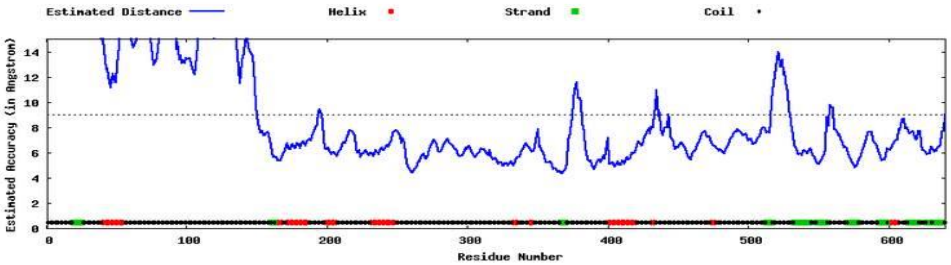
C-score = -2.83

Descargar modelo

Figura 33. Modelo tridimensional I-Tasser N°4 de la secuencia de aminoácidos de la polifenoloxidasa del lulo.

MODELO 5





C-score = -3.21

Descargar modelo

Figura 34. Modelo tridimensional I-Tasser N°5 de la secuencia de aminoácidos de la polifenoloxidasa del lulo.

Observando las figuras anteriores se puede concluir que el modelo con más índice de confianza en la predicción o con más alto C-score es el modelo número 1. En este se visualiza que los valores de distancia estimada entre los aminoacidos no son tan variables como se observa en los demás modelos tridimensionales y su valor medio que se encuentra en aproximadamente 6 angstrom es bueno, esto se debe a que entre menor sea el valor medio de la distancia entre residuos la predicción es mucho mejor, también esto se debe en que la distancia entre aminoácidos de una proteína descubierta se encuentra alrededor entre 2.5 Å y 3.5 Å[2]. Igual que en el modelado por homología, en este modelo también se tiene un botón para descargar el modelo el cual suple la misma necesidad antes mencionada. En la siguiente figura se visualizara el modelo tridimensional N°1 realizado por I-Tasser en la herramienta JMOL desde un punto de vista diferente.

Aunque la principal funcionalidad del servidor I-Tasser radica en la obtención de las estructuras terciarias de una proteína también se cuenta con la estructura secundaria como un dato fundamental para el análisis y estudio de la proteína objetivo.

[2] http://www.biorom.uma.es/contenido/av_bma/apuntes/t5/t5.htm.

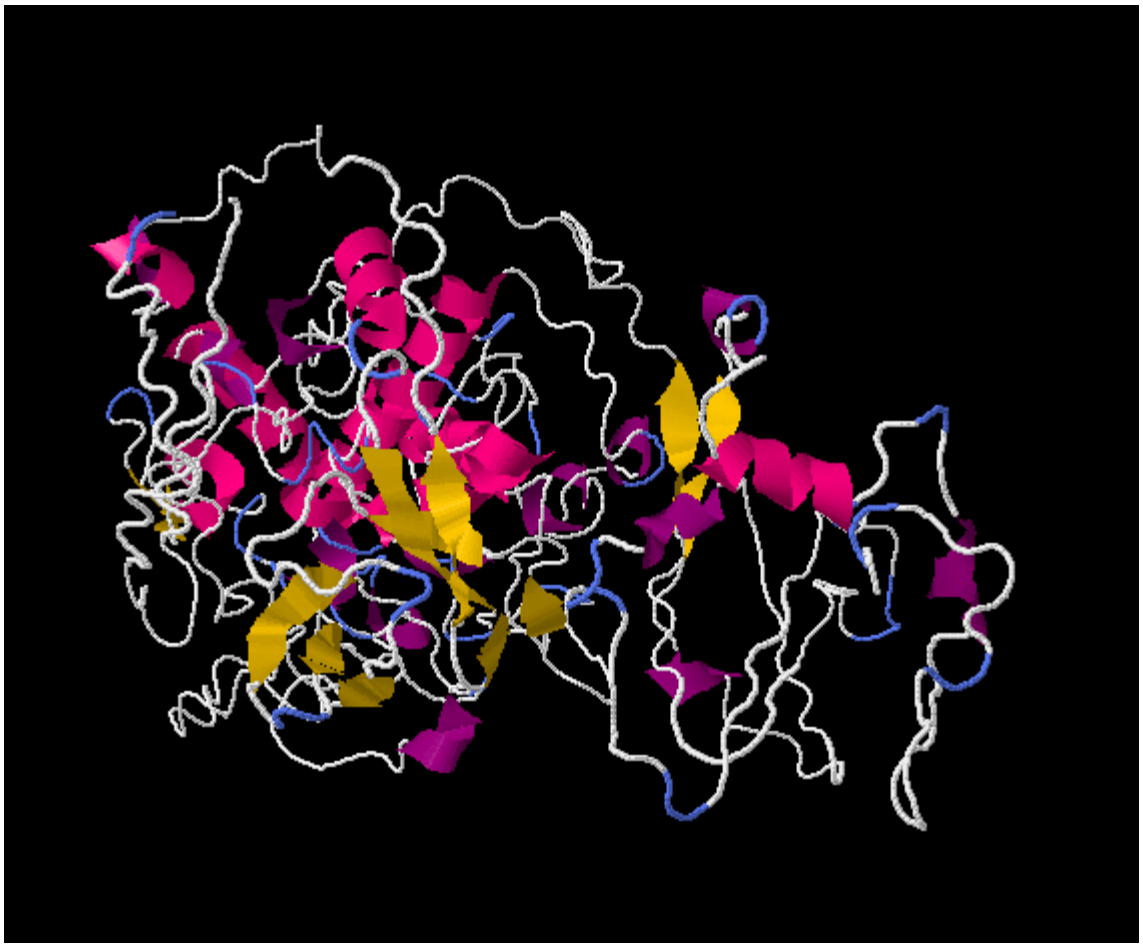


Figura 35. Modelo tridimensional I-Tasser N°1 con JMOL de la secuencia de aminoácidos de la polifenoloxidasa del lulo.

Modelado 2D JPRED4.

Con este modelado se obtiene la estructura 2D a partir una secuencia de aminoácidos, con el motivo de conocer la composición que cada residuo cumple sobre una secuencia. En este caso este servidor web entrega como respuesta lo que se observa en la figura 36.

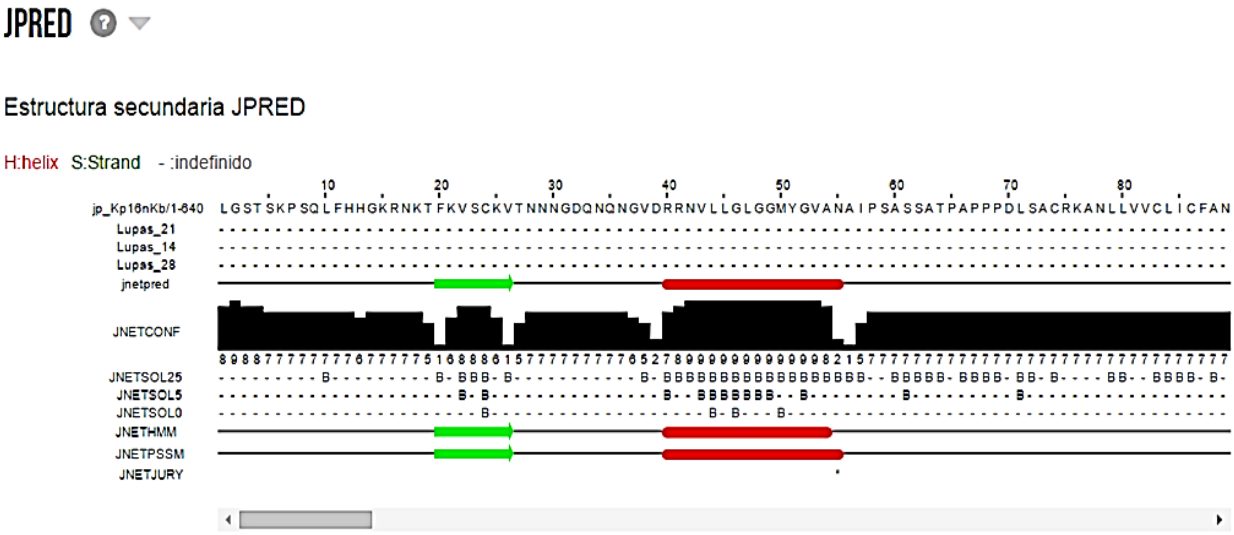


Figura 36. Modelo estructura secundaria JPRED4 de la secuencia de aminoácidos de la polifenoloxidasa del lulo.

En la figura anterior solo se puede observar una parte de la secuencia objetivo debido a la cantidad de aminoácidos que contiene. En esta imagen se observa alrededor de solo 90 aminoácidos. En este segmento se observa que la característica JNETCONF, la cual muestra que la estimación de confianza en la predicción realizada es relativamente muy buena ya que mantiene una confianza de alrededor del 70%.

En las figura 12, 29 y 36 se observan un signo de pregunta y un icono desplegable alrededor cada título del modelo. El signo de pregunta es utilizado para dar información acerca de los servidores utilizados en la predicción de este modelo, mientras el icono desplegable permite ocultar las secciones homología, ab initio y jpred.

Validación de resultados.

Con el fin de tener una referencia en cuanto al acierto de los resultados arrojados por la aplicación desarrollada se procede a validar los mismos con la información existente en la base de datos del national center of bio informatics (NCBI) y del protein data bank (PDB) comparando los mismos con los datos de una estructura determinada en laboratorio, sin embargo siendo inexistente una estructura terciaria conocida para la proteína polifenoloxidasa del lulo es imposible tener datos de comparación para la misma, por esta razón se realizó una segunda prueba con el objetivo de tomar una de las proteínas más conocidas a nivel mundial como lo es la hemoglobina específicamente una porción de la misma presente en los humanos la proteína hemoglobin beta, partial [Homo sapiens].

5.1.2. Segunda Prueba.

Para esta prueba se utiliza la secuencia de aminoacidos de la proteína hemoglobin beta, partial [Homo sapiens], esta secuencia se obtuvo de la NCBI en su formato FASTA. Esta secuencia es solo una parte pequeña de la proteína Hemoglobina. Se utilizó esta secuencia de aminoácidos para observar el funcionamiento de los diferentes servidores utilizados en la aplicación web con una secuencia con más información presente en la web con respecto a la secuencia anterior.

```
>gi|410178713|gb|AFV63187.1| hemoglobin beta, partial [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALG
```

Figura 37. Secuencia de aminoácidos de la hemoglobina beta en formato FASTA.

Se insertó la secuencia de aminoacidos en el campo que la requiere como se muestra en la figura 38, posteriormente se seleccionaron las tres opciones de predicción para predecir este secuencia. Finalmente se hizo click en el botón predecir para la obtención de los resultados.

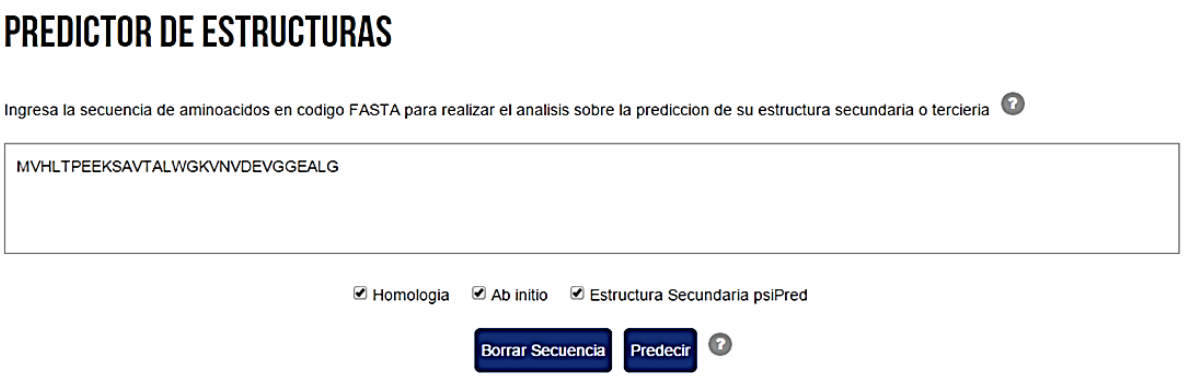


Figura 38. Inserción de la secuencia y selección de los métodos a utilizar en el aplicativo.

Al tener disponible esta secuencia para su utilización en el aplicativo web, se procede a observar los resultados arrojados que se obtuvieron por medio de los diferentes tipos de modelado presentes en la aplicación. También esta secuencia al tener alrededor de 30 aminoácidos presenta tiempos de respuesta relativamente cortos con respecto a la prueba anterior, se pasó de esperar alrededor 20 minutos para la prueba anterior, a esperar tan solo 8 minutos para esta prueba.

Modelado por Homología.

Los resultados obtenidos por medio del aplicativo web PSIBLAST se pueden observar en las siguientes figuras, posteriormente se explicarán los resultados que se obtuvieron.

Figura 39. Secuencia objetivo, secuencia más cercana a la secuencia objetivo.

Tabla de proteinas cercanas

Descripcion	Puntaje Maximo	Puntaje Total	Query Cover	Valor E	Identidad	Accession
mutant beta-globin [Homo sapiens]	96.9	96.9	100%	2e-23	100%	AAG46182.1
beta globin variant [Homo sapiens]	96.9	96.9	100%	2e-23	100%	AAP44006.1
hemoglobin beta [Homo sapiens]	96.9	96.9	100%	3e-23	100%	AFR11469.1
beta-globin thalassemia [Homo sapiens]	96.9	96.9	100%	4e-23	100%	AAA16335.1
beta-globin [Homo sapiens]	96.9	96.9	100%	4e-23	100%	AAA88069.1
truncated beta globin [Homo sapiens]	96.9	96.9	100%	5e-23	100%	ACF16769.1
beta globin [Homo sapiens]	96.9	96.9	100%	6e-23	100%	ACZ67952.1
beta globin [Homo sapiens]	96.9	96.9	100%	6e-23	100%	AAB60348.1
beta globin [Homo sapiens]	96.9	96.9	100%	2e-22	100%	AAC97372.1
hemoglobin beta chain [Homo sapiens]	96.9	96.9	100%	2e-22	100%	ADW79453.1

Figura 40. Tabla de proteínas más cercanas.

La figura 40 contiene las 10 secuencias cercanas a la secuencia objetivo. Si se desea aumentar la visualización de secuencias en la tabla, se debe proceder a hacer click sobre el icono que se encuentra en la parte superior derecha de la tabla, el cual tiene una forma redonda y un signo “+”. Al hacer click sobre icono se obtendrán las 100 secuencias más cercanas. También se observa en esta figura el valor de identidad de cada secuencia de aminoácidos con respecto a la secuencia objetivo. En comparación con la secuencia de la prueba anterior, se puede visualizar que el valor de identidad es el máximo permitido, esto se debe a que la secuencia de aminoacidos utilizados en la prueba 2 pertenece a una proteína ya conocida por lo cual la predicción debe ser del 100%, esto lo evaluaremos en la validación de resultados.

A continuación se observa el alineamiento entre la secuencia objetivo y la secuencia más cercana, como la secuencia objetivo y la más cercana tienen una similitud del 100%, el resultado del alineamiento con el servidor CLUSTAL-OMEGA indica que todas las posiciones contienen un único residuo totalmente conservado.

Figura 41. Alineamiento con CLUSTAL-OMEGA.

Para finalizar con el modelado por homología, el aplicativo inserta el resultado de la figura 42 en el servidor SWISS-MODEL, el cual entrega como resultado el estimado de calidad local, set de referencia, modelo construido y el porcentaje de acierto del modelo construido. Para este alineamiento entre las secuencias de aminoácidos este servidor entrego solo un modelo construido.

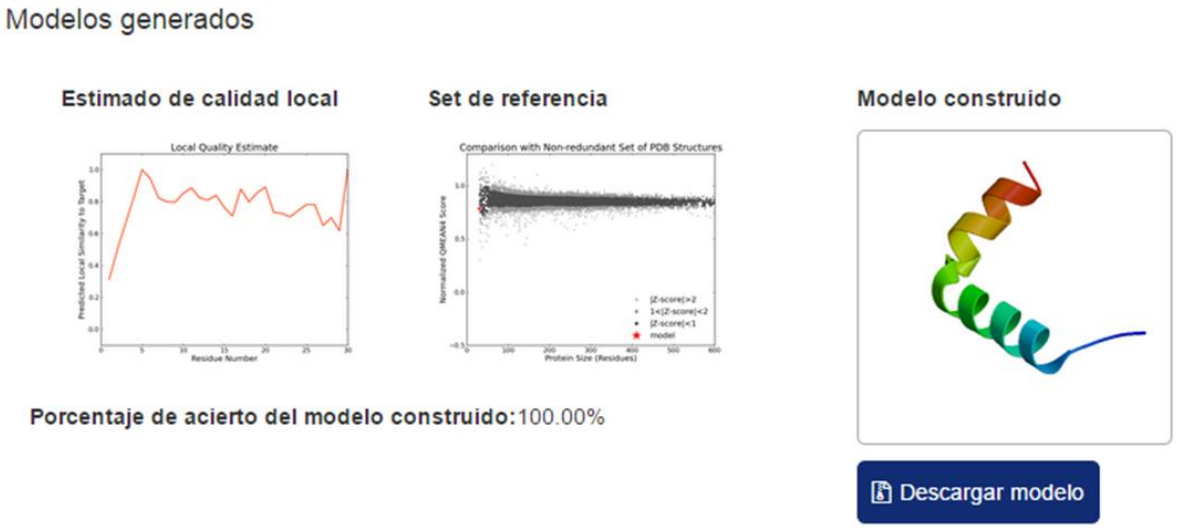


Figura 42. Modelo tridimensional SWISS-MODEL de la secuencia de aminoácidos de la hemoglobina beta.

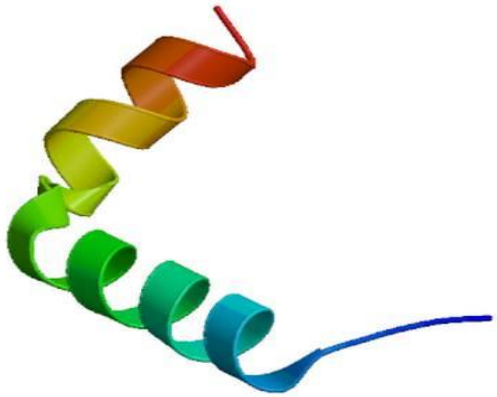


Figura 43. Modelo tridimensional SWISS-MODEL, modelo construido de la secuencia de aminoácidos de la hemoglobina beta.

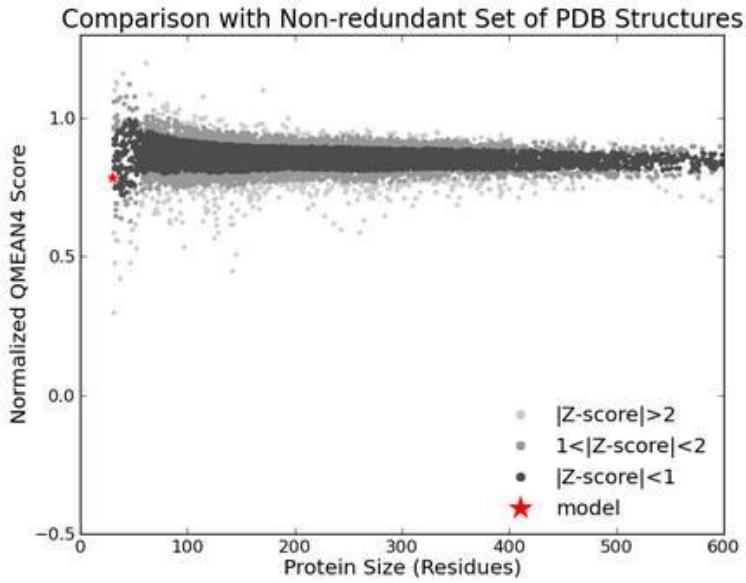


Figura 44. Modelo tridimensional SWISS-MODEL set de referencia de la secuencia de aminoácidos de la hemoglobina beta.

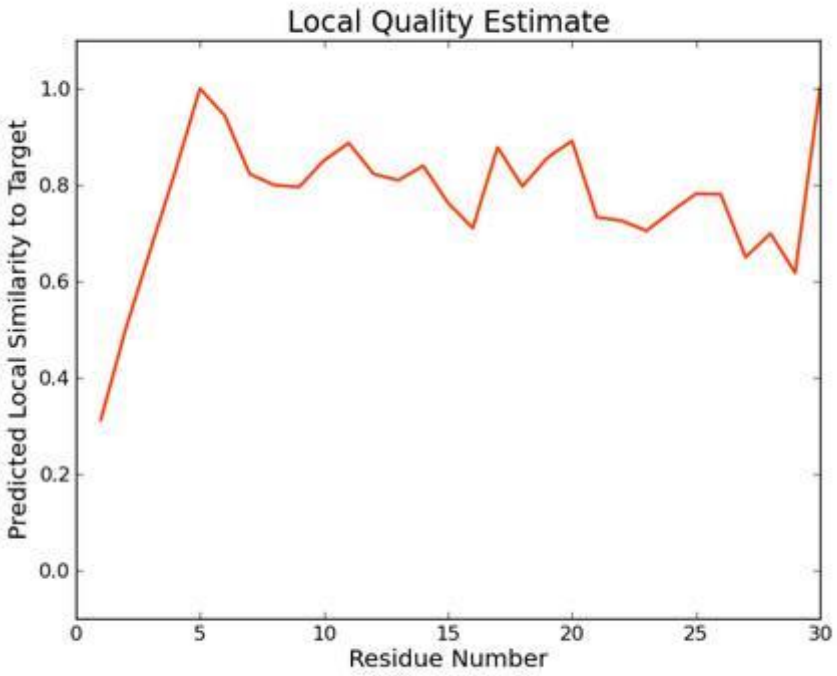


Figura 45. Modelo tridimensional SWISS-MODEL, estimado de calidad local de la secuencia de aminoácidos de la hemoglobina beta.

Como se observa en el modelo tridimensional construido a partir de su alineamiento, se concluye que el porcentaje de acierto es del 100%, este porcentaje es alto debido a la alta información que se conoce acerca de esta secuencia de aminoácidos. También se observa que el estimado de calidad local no sufre demasiada variabilidad como en los modelos tridimensionales del SWISS-MODEL en la prueba anterior.

Al descargar el modelo tridimensional dando click sobre el botón y utilizando el aplicativo JMOL se puede observar este modelo en diferentes desplazamientos.

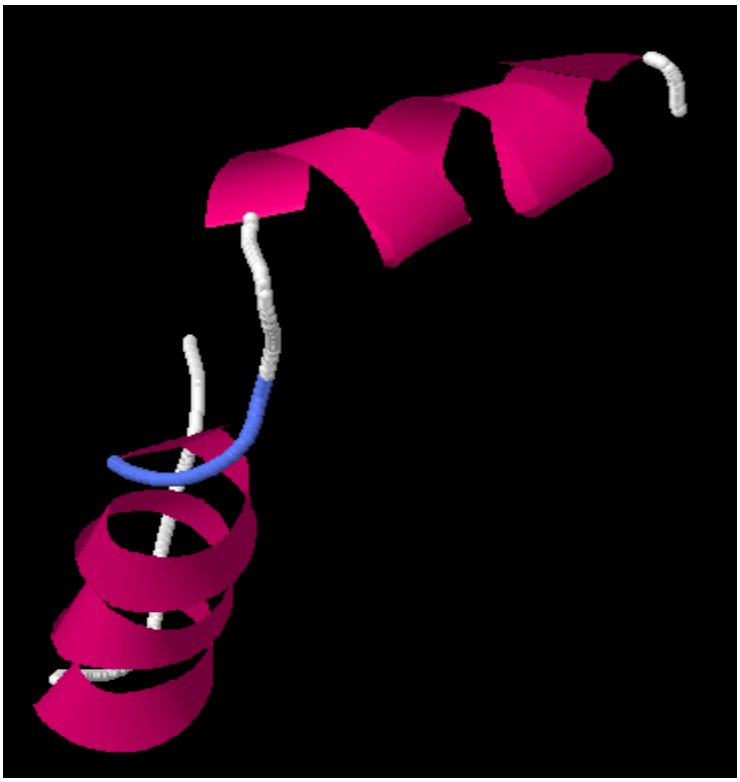


Figura 46. Modelo tridimensional SWISS-MODEL en JMOL de la secuencia de aminoácidos de la hemoglobina beta.

Modelado AB-Initio.

Para la obtención de los resultados por medio del servidor web I-TASSER, el tiempo de respuesta fue de alrededor de tan solo 5 minutos, esto es debido a su poca cantidad de aminoácidos, que para esta secuencia es de tan solo de 30 caracteres, sin embargo en algunas ocasiones debido a la cola de espera puede tomar mucho más tiempo. En la figura 47 se visualiza la estructura secundaria de la cual se puede concluir que esta secuencia solo contiene dos hélices y tres enlaces. Por lo tanto se puede concluir que el modelo tridimensional como se muestra en la figura 46 si es el correcto debido a que este presenta las mismas características que son dos hélices las cuales se encuentran en color morado y tres enlaces que son los de color blanco y azul.

AB INITIO

Estructura secundaria

```
                        20
                        |
Sequence    MVHLTPEEKSAVTALWGKVNVDEVGGEALG
Prediction  CCCCCHHHHHHHHHHHHCCCHHHCCCCCCC
Conf.Score  974778999999988735767540651139
            H:Helix; S:Strand; C:Coil
```

Figura 47. Estructura secundaria obtenido con I-TASSER de la secuencia de aminoácidos de la hemoglobina beta.

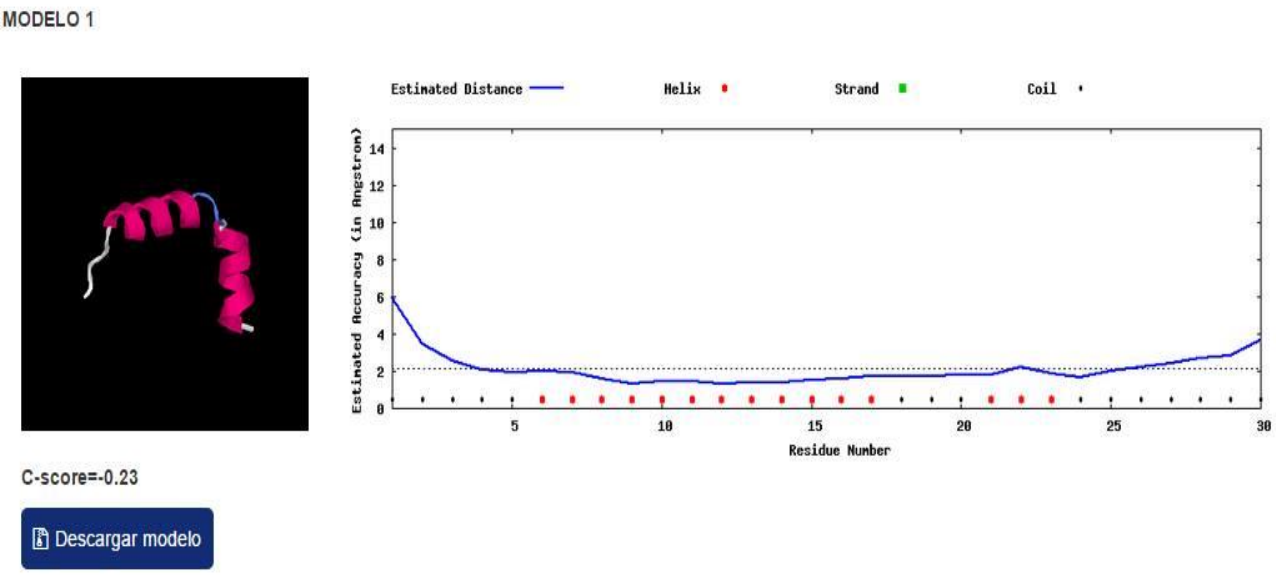
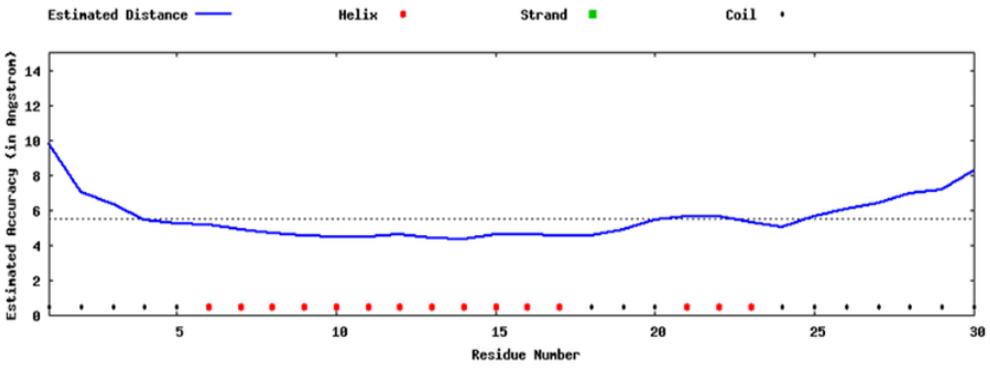


Figura 48. Modelo tridimensional I-Tasser N°1 de la secuencia de aminoácidos de la hemoglobina beta.

MODELO 2

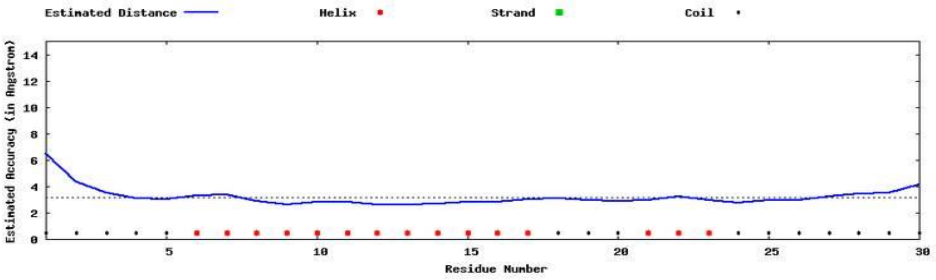


C-score = -2.39

Descargar modelo

Figura 49. Modelo tridimensional I-Tasser N°2 de la secuencia de aminoácidos de la hemoglobina beta.

MODELO 3

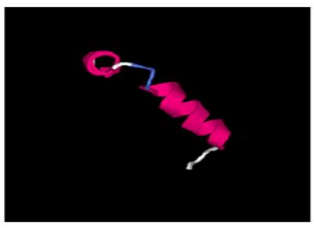


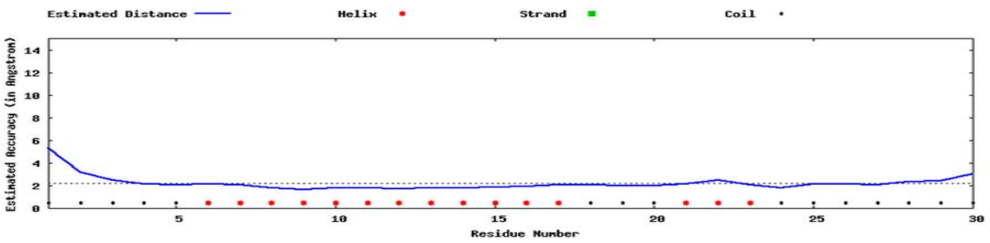
C-score = -0.81

Descargar modelo

Figura 50. Modelo tridimensional I-Tasser N°3 de la secuencia de aminoácidos de la hemoglobina beta.

MODELO 4





C-score = -0.28

Descargar modelo

Figura 51. Modelo tridimensional I-Tasser N°4 de la secuencia de aminoácidos de la hemoglobina beta.

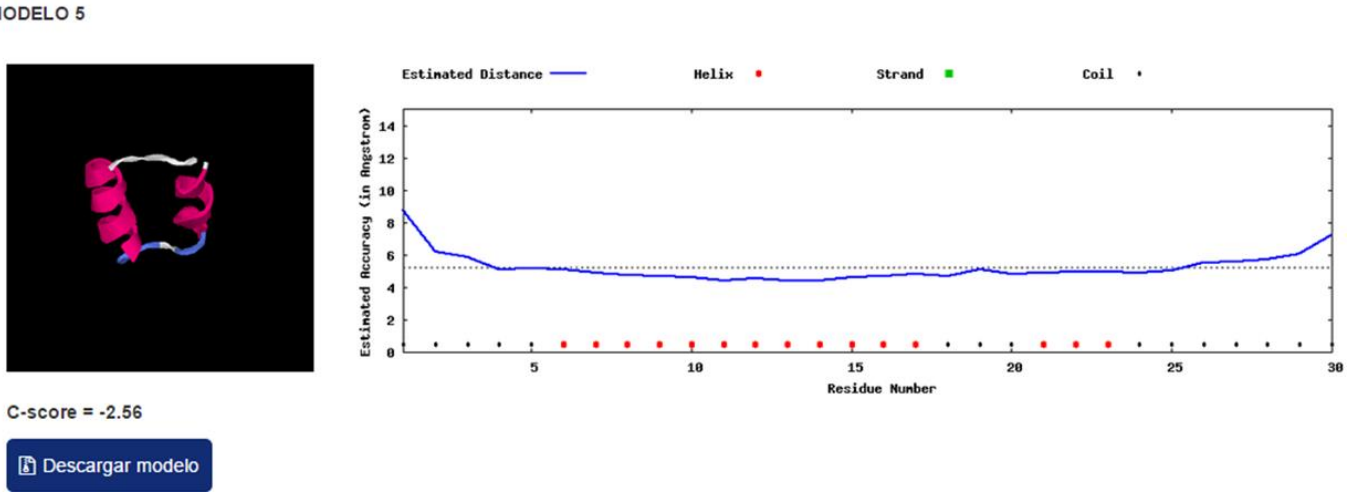


Figura 52. Modelo tridimensional I-Tasser N°5 de la secuencia de aminoácidos de la hemoglobina beta.

Observando los cinco modelos obtenidos por el método de Ab-Initio se concluye que el modelo uno siempre será el más acertado y con más alto índice de confianza en la predicción. Observando las figuras 30 y 48 se analiza que los datos entregados en la prueba dos tienes más altos índices de validez, esto se puede ver reflejado en la distancia estimada, donde para la prueba 2 su variabilidad es casi nula exceptuando algunos picos. También se puede observar que el valor medio o la distancia entre residuos es mucho menor debido a que en la primera prueba se tenía un distancia de seis angstrom mientras que en la segunda prueba se obtuvo un distancia entre residuos de dos angstrom , lo cual es beneficio para posteriores análisis biológicos.

Posteriormente de la obtención de la respuesta por Ab-Initio, se descarga el modelo tridimensional N°1 para su posterior visualización por medio del aplicativo JMOL. Esta visualización la podemos observar en la figura 53.

Figura 53. Modelo tridimensional I-Tasser N°1 con JMOL de la secuencia de aminoácidos de la hemoglobina beta.

Modelado 2D JPRED4.

Con este modelado se obtiene la estructura 2D a partir una secuencia de aminoácidos, con el motivo de conocer la composición que cada residuo cumple sobre una secuencia. En este caso este servidor web entrega como respuesta lo que se observa en la figura 54.

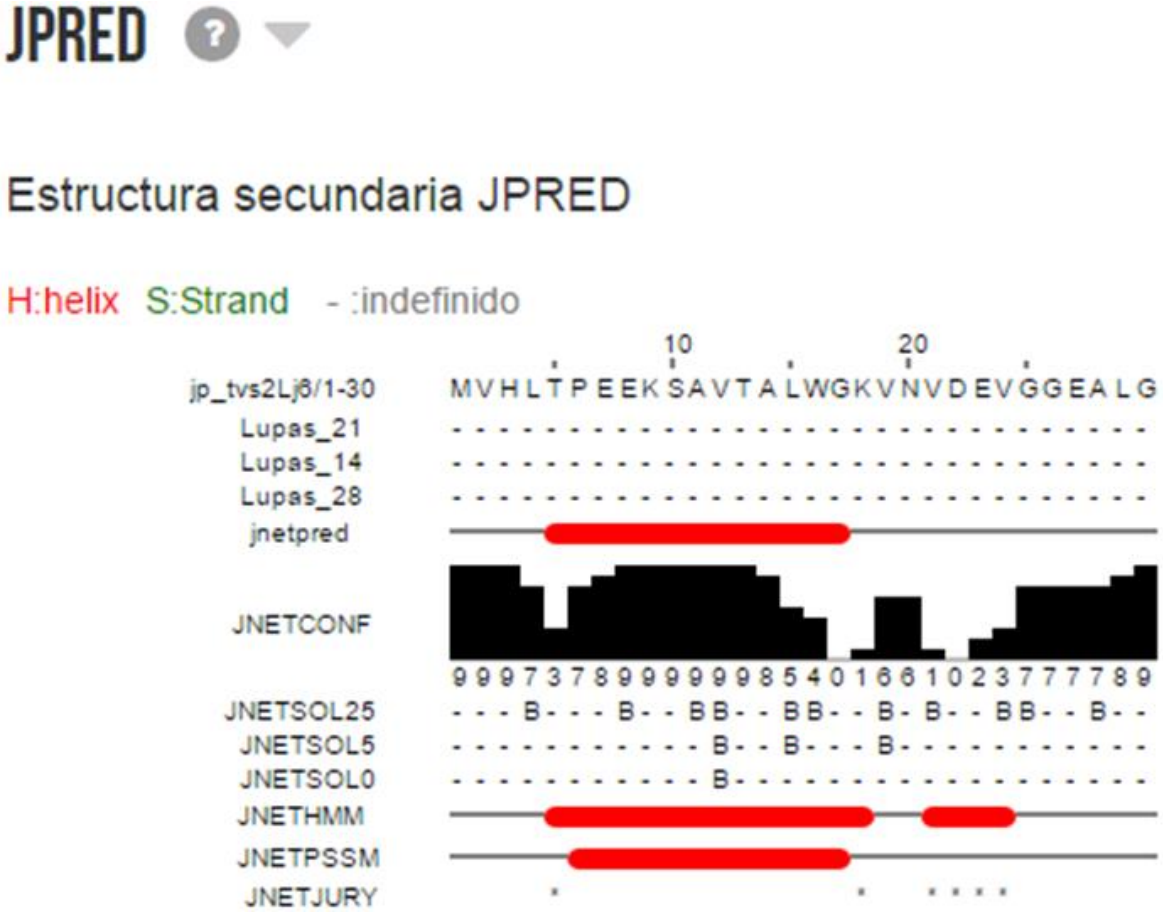


Figura 54. Modelo estructura secundaria JPRED4 de la secuencia de aminoácidos de la hemoglobina beta.

En la figura anterior se visualiza la secuencia de aminoacidos completa en la parte superior. En esta figura se observa que la característica JNETHMM, la cual muestra que tipo de residuo es una hélice o una hoja o es indefinido, se puede observar que tiene una gran similitud con la figura 47. Esto se debe a que los dos resultados tienen la misma cantidad de hélices y se encuentran entre los mismos intervalos de residuos, sin embargo el resultado ponderado jnetpred presenta una configuración un poco diferente, esto se evaluara en la validación de resultados. También de la figura anterior se observa la característica JNETCONF, la cual muestra que la estimación de confianza en la predicción realizada es elativamente muy buena ya que contiene una confianza del 62.3%.

Validación de resultados.

Para validar los resultados obtenidos mediante la aplicación realizada, se procede a comparar los resultados obtenidos mediante su uso con los datos existentes en las bases de datos del protein data bank (PDB) y del national center of bio informatics (NCBI) para la secuencia de proteína de esta prueba.

Con el fin de evaluar la estructura utilizada en la prueba, determinada como hemoglobin beta, partial [Homo sapiens] (proteína objetivo) se define la proteína que en su secuencia de aminoácidos presenta las mismas características y que ha sido evaluada experimentalmente. Debido a que la secuencia utilizada es una fracción de la secuencia completa de la proteína de la hemoglobina, se comparan los resultados con los datos de la secuencia de proteína determinada como deoxy recombinant human hemoglobin (proteína de referencia) obtenida mediante difracción de rayos x que se encuentra en la base de datos PDB. Se evaluara hasta el residuo 30 que consta de la misma secuencia determinada para esta prueba.

En la figura 55 y 56 se puede observar el alineamiento que realiza el PDB de acuerdo a la proteína objetivo y la secuencia de la proteína de referencia y el alineamiento que realiza la base de datos del NCBI.

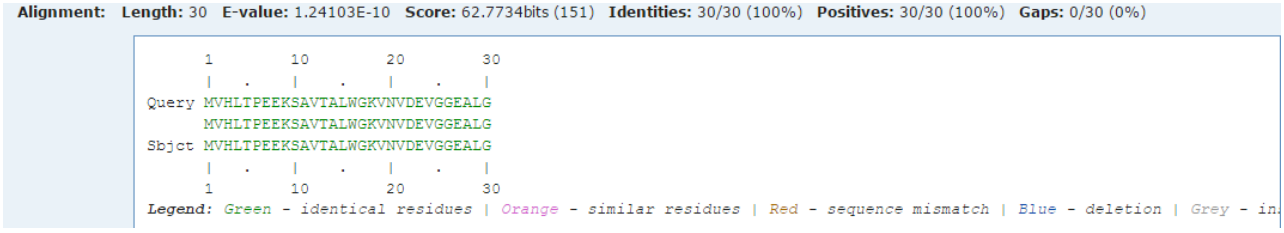


Figura 55. Alineamiento de la secuencia objetivo en el PDB.

beta-globin [Homo sapiens]

Sequence ID: gb|AAP74754.1| Length: 30 Number of Matches: 1

▶ See 3 more title(s)

Range 1: 1 to 30 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps
96.9 bits(221)	2e-23	30/30(100%)	30/30(100%)	0/30(0%)

```
Query  1   MVHLTPEEKSAVTALWGKVNVDEVGGEALG  30
           MVHLTPEEKSAVTALWGKVNVDEVGGEALG
Sbjct  1   MVHLTPEEKSAVTALWGKVNVDEVGGEALG  30
```

Figura 56. Alineamiento de la secuencia objetivo en la NCBI.

Como se puede observar en las figuras anteriores la proteína objetivo y la proteína de referencia son 100% idénticas del residuo 1 al residuo 30, por lo cual la proteína de referencia presenta las características adecuadas para validar los resultados de esta prueba, además se puede observar que los datos de la alineación obtenida mediante el método de homología por la aplicación son correctos de acuerdo a la alineación proteína objetivo- proteína cercana.

Para la estructura secundaria de la proteína de referencia, el PDB arroja lo que se muestra en la figura 57.

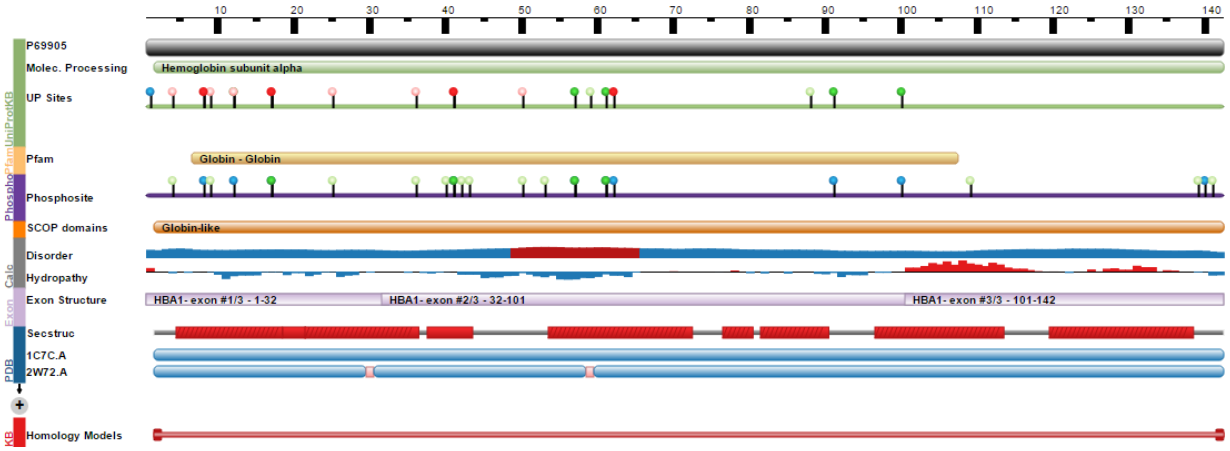


Figura 57. Estructura secundaria de la proteína de referencia en el PDB.

Este resultado indica que la estructura secundaria obtenida de forma experimental presenta el mismo comportamiento determinado por el servidor JPRED4 en su característica JNETPRED de la figura 54 hasta el residuo 30, esto indica un 100% de acierto de la herramienta de predicción de estructuras de proteínas con respecto a los resultados presentes en la base de datos del PDB.

La figura 58 muestra el resultado obtenido desde la base de datos PDB para la proteína de referencia en su estructura terciaria

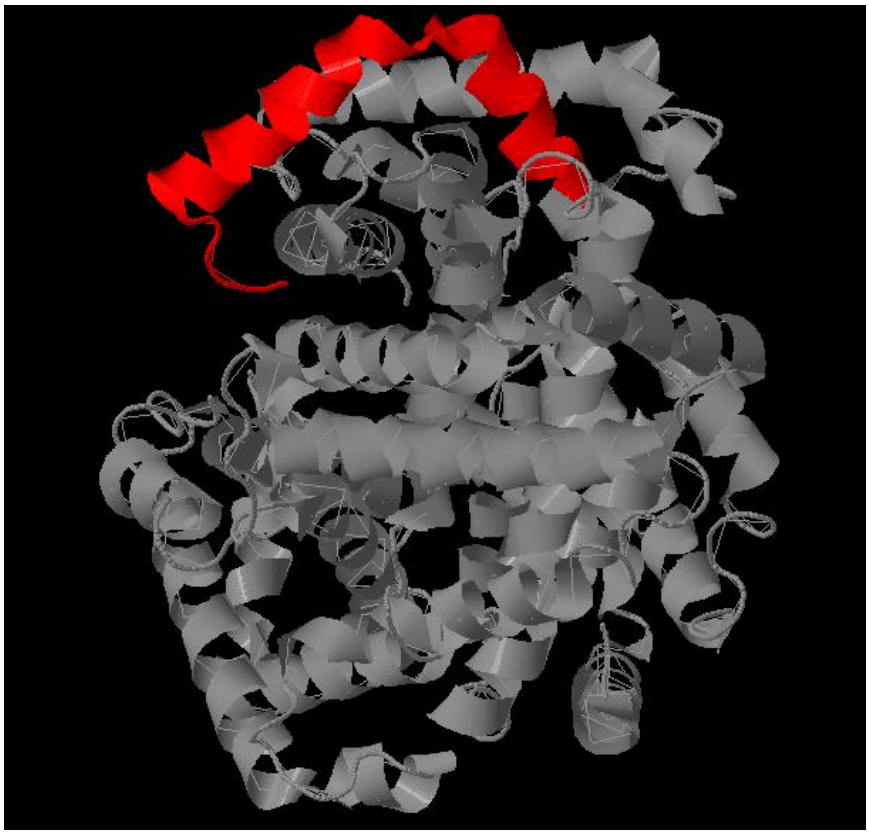


Figura 58. Alineamiento de la secuencia objetivo en el PDB.

De acuerdo a la figura anterior se puede determinar que la estructura terciaria de la proteína objetivo es idéntica a la estructura obtenida mediante los métodos de homología y ab initio. En la imagen se puede observar la cadena de 30 aminoácidos marcada en color rojo, de esta manera se puede concluir que los

resultados arrojados por la aplicación desarrollada son correctos de acuerdo a la predicción obtenida y los resultados en la base de datos PDB y NCBI.

6. CONCLUSIONES.

- Las pruebas pertinentes de funcionalidad y conectividad del aplicativo web desarrollado fueron exitosas, evidenciando los resultados esperados en cuanto a conexión con las bases de datos, servidores de predicción y visualización de la información requerida.
- Los resultados obtenidos por medio del aplicativo web con respecto a los resultados entregados por bases de datos con información ya determinada como la base de datos NCBI (National Center for Biotechnology Information) y PDB (Protein Data Bank) fueron satisfactorios como predicciones adecuadas.
- Los métodos de predicción de estructuras de proteínas presentan resultados muy diversos de acuerdo al tipo de estructura de proteínas a predecir.
- Los porcentajes de acierto y puntajes de predicción dependen de la cantidad de información acerca de diferentes tipos de proteínas presentes en las bases de datos de proteínas.
- Los algoritmos de automatización y de control del web driving presentaron tiempos de ejecución muy cortos por lo cual el tiempo de predicción depende en gran parte de la respuesta de los diferentes servidores de predicción.
- Al almacenar la información generada por la predicción de diferentes tipos de proteínas en una base de datos se disminuyen considerablemente los tiempos de respuesta de la aplicación al predecir las proteínas ya determinadas, además de facilitar la recopilación de datos para su posterior uso.

BIBLIOGRAFÍA

[1] Zhang, Yang. 2008. Progress and challenges in protein structure prediction.

[2] Petrey, Donald, y Xiang, Zhexin, y Gimpelev, Marina. 15 octubre 2003. Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. Proteins: Structure, Function, and Bioinformatics, no. 53: 430-435.

[3] Segui. Matilde Julián. Estructura y Propiedades de las proteínas.

[4] California, U. (s.f.). CASP. Recuperado el 2014, de Protein Structure Prediction Center: http://predictioncenter.org/

[5] Qian, Yaorong. 1993. «Kinetics of peptide hydrolysis and amino acid decomposition at high temperature». Geochimica et Cosmochimica.

[6] Peretó, Julio. 2007. Fundamentos de Bioquimica.

[7] Segui. Matilde Julián. Estructura y Propiedades de las proteínas.

[8] Donald Voet, y Judith G. Voet, y Charlotte W. Pratt. 2007. Fundamentos de Bioquimica.

[9] Ramachandran, S., Dokholyan, N. 2012. Homology Modeling: Generating Structural Models to Understand Protein Function and Mechanism.

[10] Zhang, Yang. 2008. Progress and challenges in protein structure prediction.

[11] Cobo, Ángel, y Gómez, Patricia, y Pérez, Daniel, y Rocha, Roció. 2005. PHP y MySQL: Tecnología para el desarrollo de aplicaciones WEB. Ediciones Díaz de Santos.

[12] Bassi, Sebastian. Python en 8 clases: Aprendiendo a programar con Python.

[13] Beltrán, Aries. 2013. Chapter 1. Getting Started. Getting Started whith PhantomJS.

[14] Medicine, N. L. (s.f.). Ncbi. Recuperado Abril de 2015, de Psi-Blast: http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&PAGE=Proteins&PROGRAM=blastp&RUN_PSIBLAST=on

[15] EBI. (2014). Clustal Omega. Recuperado Abril de 2015, de http://www.ebi.ac.uk/Tools/msa/clustalo/

[16] Bassel, U. (s.f.). *Biozentrum*. Recuperado Abril de 2015, de SWISS-MODEL: http://swissmodel.expasy.org/interactive

[17] A Roy, A Kucukural, Y Zhang. I-TASSER: a unified platform for automated protein structure and function prediction. Nature Protocols, 5: 725-738 (2010).

[18] Mimouni, Naila, y Lunter, Gerton, y Deane, Charlotte. Hidden Markov Models for Protein Sequence Alignment. 3-4.

[19] NCBI. (s.f.). PSSM. Recuperado el 02 de 05 de 2015, de http://www.ncbi.nlm.nih.gov/Class/Structure/pssm/pssm_viewer.cgi

[20] Universidad de Basel. (s.f.). Introduction to SWISS-MODEL Workspace. Recuperado el 11 de Mayo de 2015, de http://swissmodel.expasy.org/docs/help

[21] Benkert, P., Biasini, M. and Schwede, T. (2011). "Toward the estimation of the absolute quality of individual protein structure models." Bioinformatics (2010). doi: 10.1093/bioinformatics/btq662