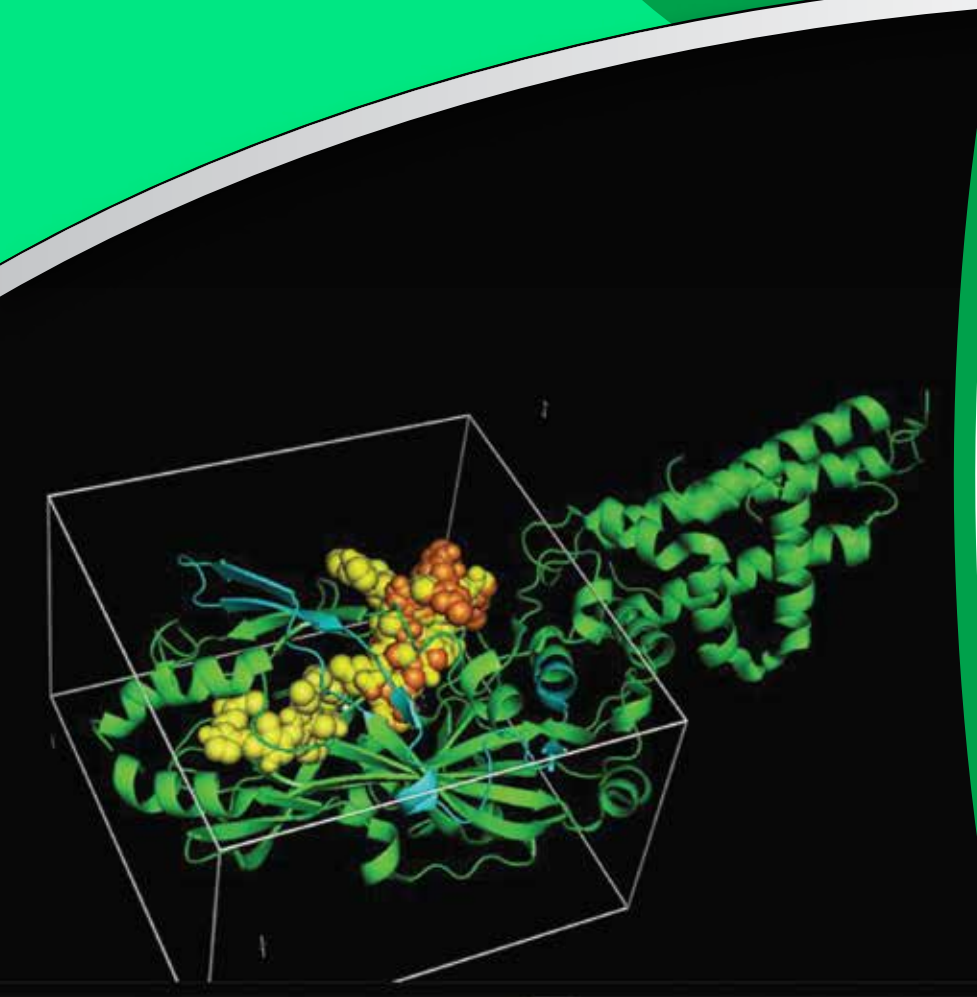
UNIVERSIDAD AUTÓNOMA METROPOLITANA
Casa abierta al tiempo
Unidad Iztapalapa

Bioinformática: un enfoque en el Proteoma



Carolina **Campos Muñiz**

Elizabeth **Hernández Pérez**

Iris Natzielly **Serratos Álvarez**

cbs
División de Ciencias Biológicas y de la Salud

Bioinformática: un enfoque en el Proteoma

Carolina **Campos Muñiz**

Elizabeth **Hernández Pérez**

Iris Natzielly
Serratos Álvarez

UNIVERSIDAD AUTÓNOMA METROPOLITANA UNIDAD IZTAPALAPA

RECTOR GENERAL
Dr. José Antonio De Los Reyes Heredia

SECRETARIA GENERAL
Norma Rondero López

UNIDAD IZTAPALAPA

RECTOR
Verónica Medina Bañuelos

SECRETARIO
Juan José Ambriz García

DIRECTOR DE LA DIVISIÓN DE C.B.S.
José Luis Gómez Olivares

COORDINADOR DE EXTENSIÓN UNIVERSITARIA
Mtro. Federico Bañuelos Bárcena

JEFE DE LA SECCIÓN DE PRODUCCIÓN EDITORIAL
Lic. Adrián Felipe Valencia Llamas

Primera edición 2022

ISBN: 978-607-28-2726-4

UNIVERSIDAD AUTÓNOMA METROPOLITANA UNIDAD IZTAPALAPA

Av. Ferrocarril San Rafael Atlixco, Núm. 186, Col. Leyes de Reforma 1 A Sección,
Alcaldía Iztapalapa, C.P. 09310, Ciudad de México

Impreso y hecho en México/Printed in Mexico

Índice

Introducción

Este manual pretende que los alumnos adquieran una visión general sobre las aplicaciones bioinformáticas comúnmente utilizadas en el estudio de las proteínas y los fundamentos que las soportan. El objetivo esencial es que los alumnos utilicen estas herramientas para el estudio de estructuras biológicas, modelado de proteínas y simulación de procesos, sirviendo como apoyo a los planes de estudio en el área de ciencias biológicas. Se introduce al manejo de las herramientas bioinformáticas mediante prácticas dirigidas con ejemplos concretos, contemplando un abanico de bases de datos para el análisis del proteoma. El manual consta de 13 prácticas con material gráfico que guiarán paso a paso y de manera eficiente para un aprendizaje significativo.

Un recurso para el estudio del proteoma es conocer y manejar las herramientas que parten desde el almacenamiento de datos hasta el modelaje de proteínas que involucra la predicción de sus estructuras, funciones y localización celular. Cabe destacar que, gracias a los avances de la bioinformática se han diseñado nuevas macromoléculas para el desarrollo de nuevos fármacos, lo que ha permitido el avance en el estudio de diferentes áreas de investigación como la biotecnología, la biomedicina, la biofísica, entre otros.

Para resolver las diversas problemáticas de las disciplinas científicas desde la perspectiva bioinformática se requieren bases de datos biológicas, indispensables para comprender y explicar algunos fenómenos, desde la estructura biomolecular y su interacción, el metabolismo de los organismos de interés y su evolución. Por otro lado, las herramientas bioinformáticas ayudarán al diagnóstico de patologías, al desarrollo de medicamentos con el fin de diseñar terapias contra algunas enfermedades, así como a entender las relaciones entre organismos.

Agradecimientos

Las autoras agradecen su valiosa aportación a los alumnos Jimena Rodríguez Carbo, Luis Ángel Carrasco Sánchez y Esteban Rafael Ramírez Pérez por el soporte técnico para realizar las prácticas.

Práctica 1

Uso de las base de datos bioinformáticas

NCBI

Introducción

Según el *National Center for Biotecnology Information (NCBI)*, la bioinformática "consiste en el acercamiento computacional al manejo y análisis de la información biomédica". La bioinformática se ha convertido en una herramienta importante de la investigación académica a nivel de licenciatura y posgrado.

Uno de los principales objetivos de las bases de datos biológicos es almacenar información, organizar y compartirla de una manera estructurada que permita una búsqueda eficiente. Otro fin es facilitar la visualización y recuperación de datos a través de aplicaciones de computación para el intercambio y la integración de la información de una manera automatizada. Las bases de datos incrementan y se actualizan día con día mismas que se encuentran contenidas principalmente en las siguientes páginas:

- https://www.NCBI.nlm.nih.gov/protein/
- https://www.rcsb.org

Estas bases de datos se almacenan y organizan en dos bancos de datos principales:

El *GenBank* en el *National Institutes of Health* (*NIH*), Bethesda, y el *EMBL Sequence Data Base* en el laboratorio de biología molecular europeo en Heidelberg, Alemania. Estas bases de datos se enriquecen continuamente con nuevas secuencias identificadas y se encuentran a disposición en Internet. Es importante mencionar al líder de los suministradores de información: el NCBI, que fue fundado en 1988 como una división de la biblioteca nacional de medicina en Estados Unidos y está ubicada en el campus de los Institutos Nacionales de Salud (*NIH*).

El objetivo del *NCBI* es proporcionar información biológica además de ser una herramienta importante como apoyo para estudiar los procesos moleculares y genéticos subyacentes a la investigación básica y aplicada en ciencias biológicas. Los objetivos específicos incluyen la creación de sistemas automáticos para almacenar y analizar la información, el desarrollo de métodos avanzados para el procesamiento de esta con computadoras para facilitar el acceso a los usuarios a las bases de datos y los programas, además de la coordinación de esfuerzos para reunir información biotecnológica de todo el mundo.

Objetivo general

- Conocer algunas de las bases de datos bioinformáticas disponibles en Internet con el fin de identificar las publicaciones sobre temas de interés biológico.

Objetivos particulares

1. Conocer diferentes bases de datos bioinformáticas disponibles en Internet con el objeto de obtener información acerca de las proteínas.
2. Manejar los buscadores de las bases de datos para obtener información sobre las proteínas de interés.

Requerimientospara la elaboración de la práctica

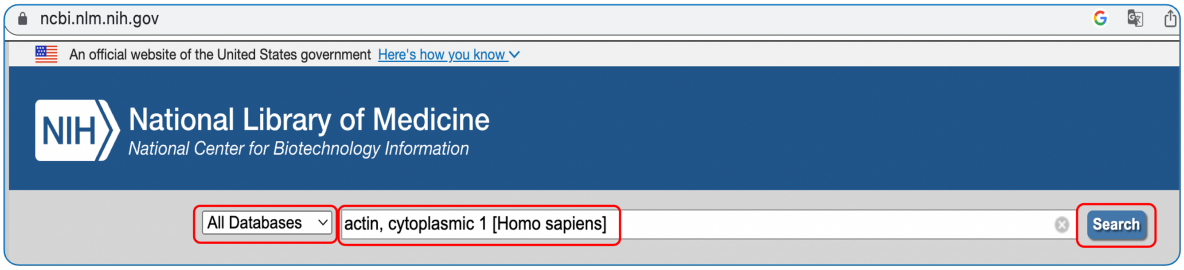
- Dispositivo con conexión a Internet.

Nota: No es necesario descargar ningún programa para realizar esta práctica, se trabaja en línea.

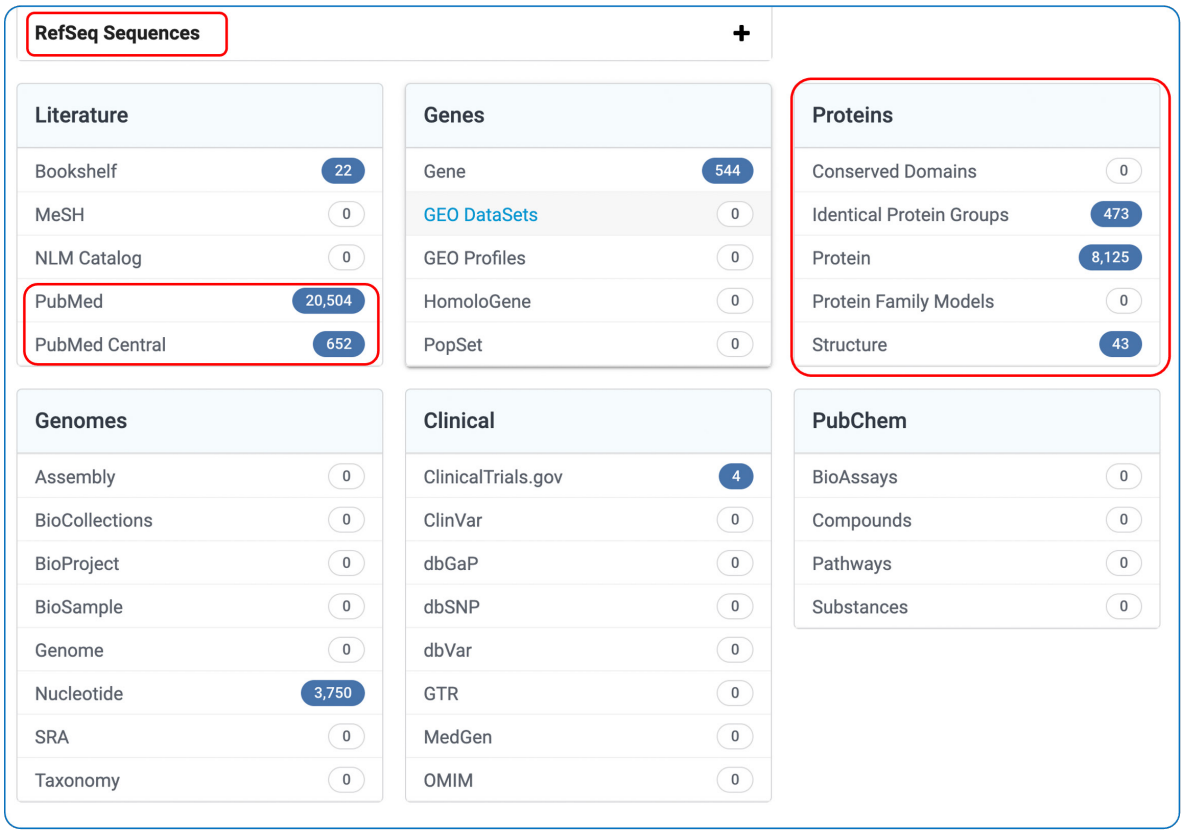
Procedimiento

1. Acceder a la base de datos de *NCBI* a través del siguiente enlace (*link*) https://www.NCBI.nlm.nih.gov/

2. Introducir el nombre de la proteína de interés, por ejemplo. Actina (*actin, cytoplasmic* 1 [*Homo sapiens*]).



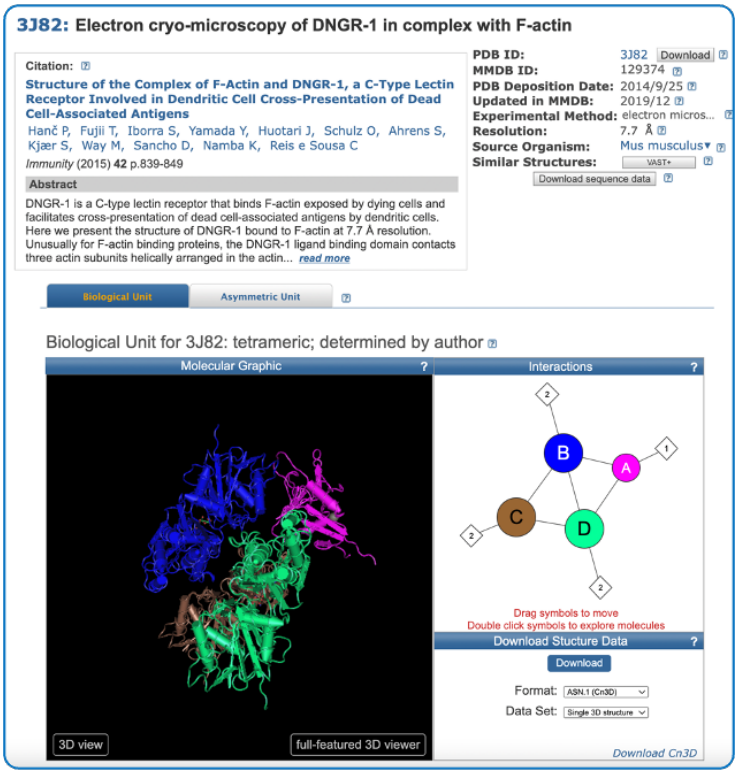
3. Al dar clic en buscar (*"Search"*) se desplegará el siguiente menú:



4. Explorar los resultados de la base de datos. En ella se encuentran publicaciones sobre la proteína: como estructura, secuencia de nucleótidos y de aminoácidos, etc.

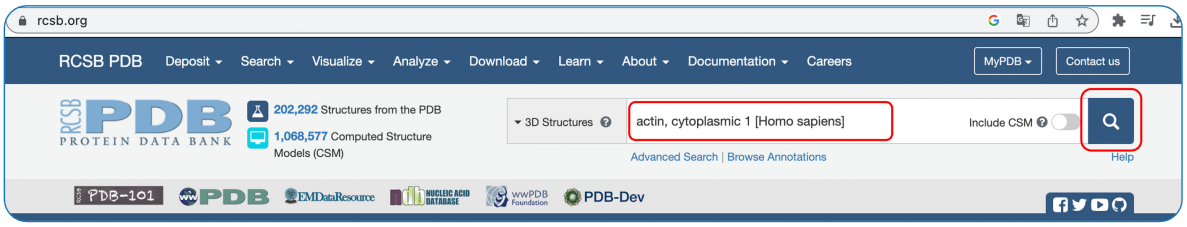
5. Acceder a *PubMed* e investigar de manera general sobre el contenido de los artículos que involucran a la proteína.

6. Acceder a la información sobre la estructura de la proteína ("*Structure*") de interés y a partir del resultado obtener la siguiente información: estructura 3D de la proteína, método empleado para determinar la estructura y la resolución atómica.



7. Acceder a la página de *Protein Data Bank* a partir del siguiente enlace: https://www.rcsb.org/

8. Introducir el nombre de la misma proteína y a partir de los resultados comparar la información obtenida de ambas bases de datos.



Cuestionario y/o ejercicios complementarios

1. ¿Cuántos artículos aparecen en *PubMed* de elastina (*elastin*) como término de búsqueda? En la página se encuentra una serie de opciones que permiten realizar una búsqueda más específica mediante opciones de filtros (*NCBI Filters*). Filtre los resultados obtenidos para conocer cuántos artículos han sido publicados en los últimos 5 años ¿Cuál es el resultado?

2. Filtrar los resultados de modo que únicamente aparezcan aquellos artículos que sean revisiones. ¿Cuántos artículos aparecen después de filtrar los resultados?

3. Realizar la búsqueda de elastina y *elastin* en Google Académico (https://scholar.google.com/) ¿Cuál es el resultado de la búsqueda?

4. Comparar los resultados obtenidos en *PubMed* y Google Académico.

5. ¿Qué motor de búsqueda es más eficiente?

Bibliografía

- Atwood, T., y Parry-Smith, D. (2002). *Introducción a la bioinformática*. Madrid, España: Pearson Educación, S.A.
- Capel, J., y Yuste, F. (2016). *Manual de prácticas de bioinformática*. Andalucía, España: Almería.
- Lodish, H., Berk, A., Kaiser, C., Kriegeer, M., Bretscher, A., Ploegh, H., Amon, A., y Scott, M. (2016). *Biología celular y molecular*. Buenos Aires, Argentina: Médica Panamericana.
- National Center for Biotechnology Information (NCBI)(Internet). Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; (1988) – [citado el 30 de octubre del 2019]. Disponible en: https://www.NCBI.nlm.nih.gov/

Práctica 2

Comparación de secuencias de aminoácidos

BLAST

Introducción

Los 20 aminoácidos (aa) esenciales que conforman las proteínas se encuentran unidos por enlaces peptídicos, en diferente número y posición dependiendo de la proteína. La secuencia de los aa enlazados contiene la información que se necesita para generar una molécula proteica con la forma tridimensional única que determina la función.

La complejidad de la estructura proteica se analiza mejor al considerar la molécula en términos de su organización, como la estructura primaria de forma lineal que no incluye ninguna fuerza o enlace y está determinada por la secuencia de aa en la cadena proteica, es decir, la especificación del número de aa de cada clase, y del orden en que están alineados. Por ejemplo, al comparar la secuencia de aa de la proteína codificada por un gen con las secuencias de proteínas con función conocida se pueden buscar similitudes de secuencias que proporcionen datos para conocer la posible función de la proteína. Debido a la degeneración en el código genético, las proteínas relacionadas invariablemente exhibirán más similitudes en su secuencia de aa que en la secuencia de nucleótidos de los genes que las codifican. Por esta razón, se suelen comparar las secuencias de aa en lugar de las secuencias de DNA correspondientes.

La búsqueda de similitud de secuencias, típicamente con *BLAST*, es la estrategia más utilizada y confiable para caracterizar secuencias recién determinadas, se pueden identificar proteínas o genes homólogos al detectar una similitud estadísticamente significativa misma que refleja un ancestro común. En sentido estricto, la homología se refiere solamente a proteínas que han evolucionado a partir del mismo gen o que provienen de un gen primitivo común. En los casos de proteínas estrechamente relacionadas es fácil detectar la homología gracias a la elevada similitud de secuencia.

Para comparar las secuencias se alinean una debajo de otra considerando el código de una letra (Tabla 1). Si en ambas secuencias se encuentra un aa en idéntica posición, se dice que existe una coincidencia o match; si los aa son diferentes se utiliza el término mismatch. Ocasionalmente hay también brechas o gaps, es decir, posiciones en una secuencia para las cuales no se encuentra correspondencia. Los gaps corresponden a aa insertados o eliminados, que pueden ser mutaciones por inserciones o deleciones que tuvieron lugar en el transcurso de la evolución. Si se construyen las relaciones de matches para un número común de aa en la secuencia ordenada, se consigue una medida cuantitativa para la identidad de secuencia. Sin embargo, no solo se evalúa la identidad, sino también la similitud de las correspondientes posiciones de los aa.

Tabla 1. Códigos de los aminoácidos de una y tres letras.

Aminoácido	Código de tres letras	*FASTA*	Aminoácido	Código de tres letras	*FASTA*
Ác. Aspártico	Asp	D	Isoleucina	Ile	I
Ác. Glutámico	Glu	E	Leucina	Leu	L
Alanina	Ala	A	Lisina	Lys	K
Arginina	Arg	R	Metionina	Met	M
Asparagina	Asn	N	Prolina	Pro	P
Cisteína	Cys	C	Serina	Ser	S
Fenilalanina	Phe	F	Tirosina	Tyr	Y
Glicina	Gly	G	Treonina	Thr	T
Glutamina	Gln	Q	Triptófano	Trp	W
Histidina	His	H	Valina	Val	V

BLAST

El algoritmo *BLAST* de computación utilizado con mayor frecuencia para la comparación de secuencias se conoce como *BLAST* (*Basic Local Aligment Search Tool*). *BLAST* divide la secuencia de la nueva proteína conocida como la secuencia consultada (*query*), en fragmentos más cortos y luego busca en la base de datos coincidencias significativas con cualquiera de las secuencias almacenadas. Este programa de coincidencias asigna una puntuación elevada a los aminoácidos con idénticas coincidencias y una puntuación baja entre aminoácidos que están relacionados, pero no idénticos, por ejemplo, hidrófobos, polares, cargados positivamente, cargados negativamente.

Cuando se encuentra una coincidencia importante para un segmento, el algoritmo *BLAST* buscará localmente la región de similitud. Después de que se haya completado la búsqueda, el programa clasifica las coincidencias entre la proteína bajo estudio y las diversas proteínas conocidas con base en sus *valores p*. Este parámetro es una medida de probabilidad de encontrar tal grado de similitud entre dos secuencias de proteínas por azar. Mientras más bajo es el valor *p*, mayor es la similitud entre dos secuencias. Un valor *p* menor alrededor de 10^{-3} suele ser considerado como una evidencia significativa de que dos proteínas comparten un ancestro común. Esta comparación de secuencias es una herramienta bioinformática básica que permite extraer información funcional, estructural y evolutiva contenida en las secuencias biológicas.

Objetivo general

- Comparar dos o más secuencias de proteínas mediante el algoritmo *BLAST* para conocer su grado de similitud.

Objetivos particulares

1. Contar con elementos teóricos para analizar y diferenciar la identidad, la homología o similitud de proteínas y predecir la estructura y función.

2. Conocer la homología o similitud entre secuencias de las proteínas mediante su alineamiento para predecir su posible función.

Requerimientos para la elaboración de la práctica

- Dispositivo con conexión a Internet.

Nota: No es necesario descargar ningún programa para realizar esta práctica, se trabaja en línea.

Procedimiento

Esta práctica consta de tres secciones:

I. Obtención del formato *FASTA* de una proteína en *NCBI*.

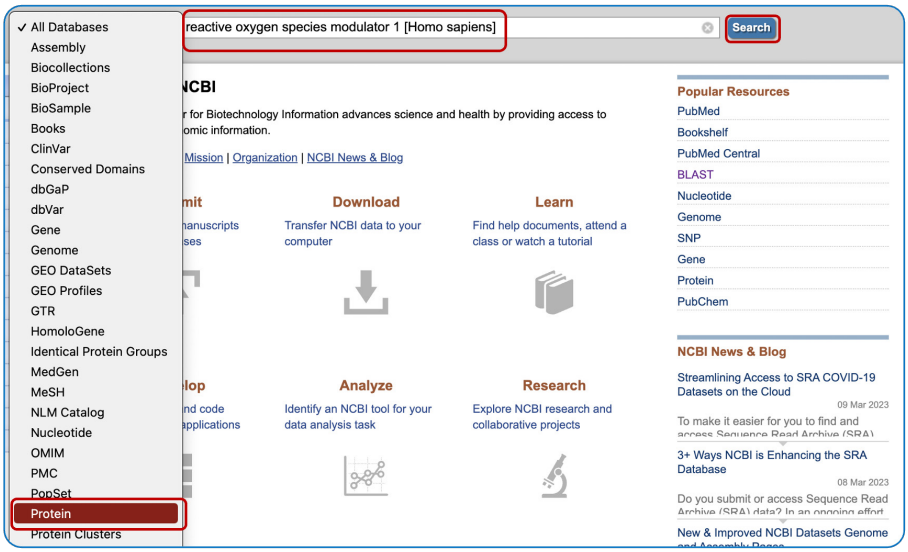
II. Alineamiento de dos secuencias.

III. Alineamiento de secuencias múltiples.

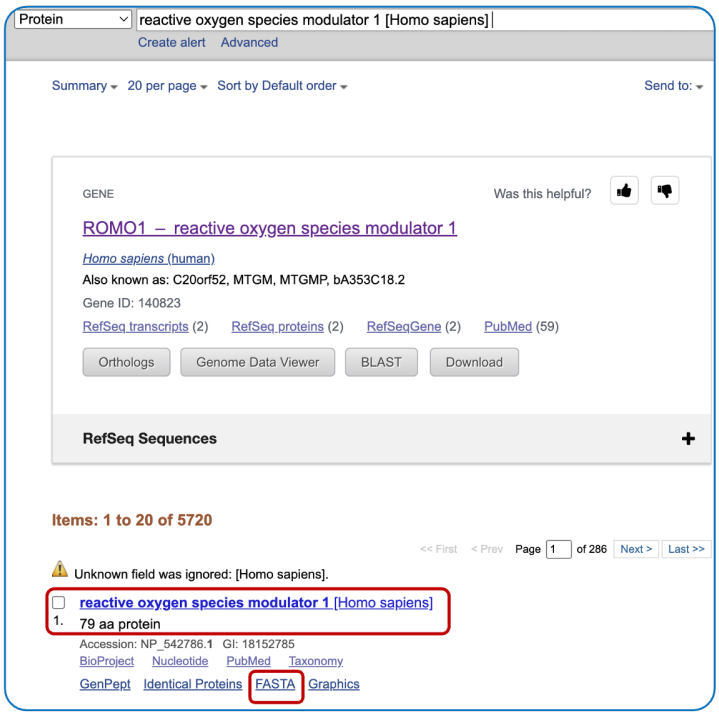
I. Obtención del formato *FASTA* de una proteína en *NCBI*

El formato *FASTA* tiene un encabezado que comienza con el signo ">" en donde se tiene la descripción de la proteína a la que corresponde. El siguiente renglón contiene la secuencia de aa en formato de una letra, la estructura primaria de la proteína (Tabla 1).

1. Para poder realizar el alineamiento de las secuencias proteicas se requiere obtener la secuencia de la proteína de interés en formato *FASTA*, para ello se debe acceder a *NCBI* (https://www.NCBI.nlm.nih.gov/), en el botón *"All Databases"* elegir la pestaña "*Protein*", en la ventana de búsqueda se coloca el nombre de la proteína, por ejemplo, *Romo*1, [*Homo sapiens*]. Dar clic en "*Search*".

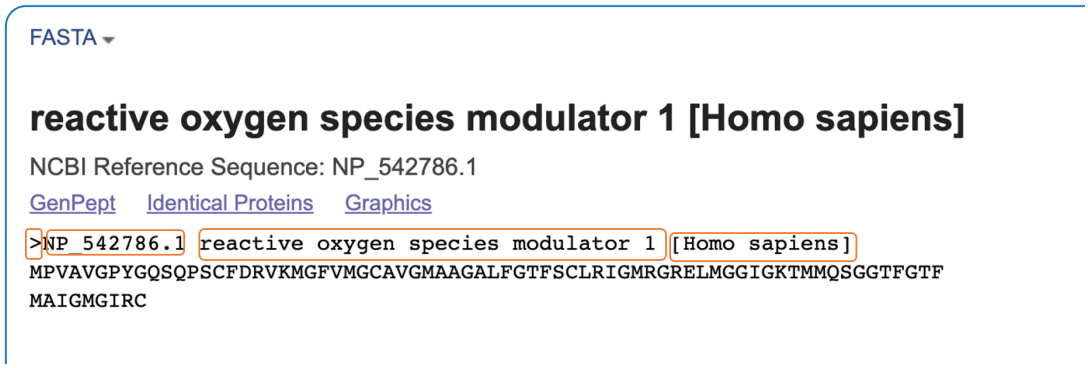


2. Dar clic en *FASTA* para obtener la secuencia.



3. Se despliega la información de la proteína, en donde la primera línea muestra la definición del formato *FASTA*, la cual se caracteriza por comenzar con el símbolo de mayor qué ">", seguido de un nombre y un identificador único ("*Accession*") de la proteína seguido por la secuencia en el formato *FASTA*.

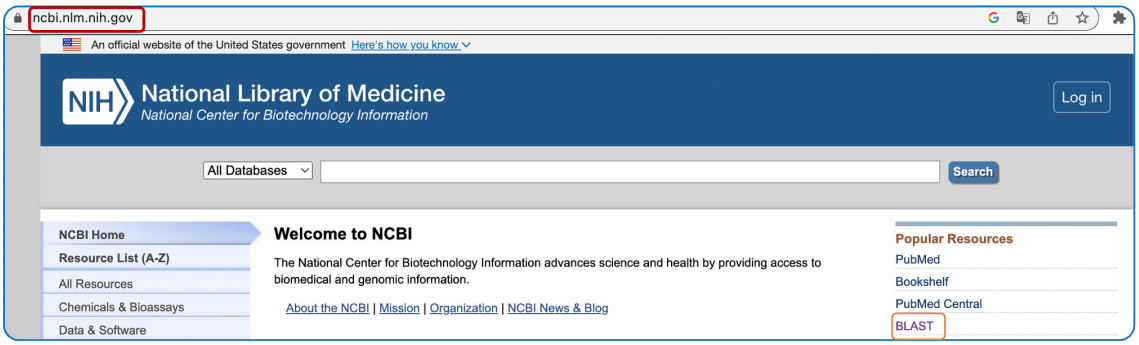
4. De la misma forma obtener el formato *FASTA* para *paralemmin*-1 *isoform* 1 [*Mus musculus*].



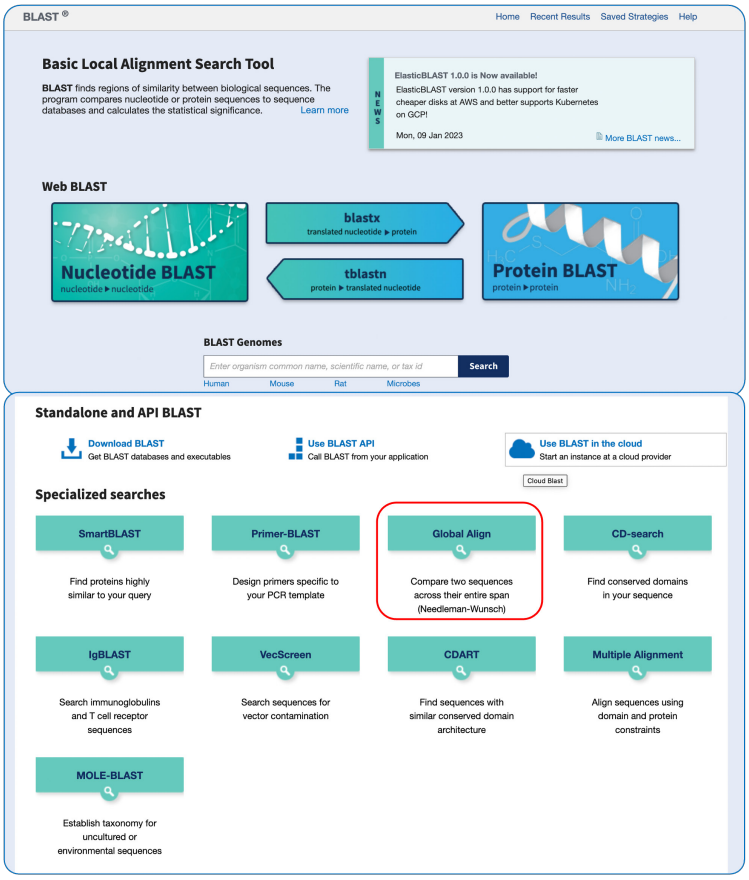
II. Alineamiento de dos secuencias

El poder del análisis de las secuencias reside en la capacidad de tomar simultáneamente secuencias relacionadas y expresar el grado de similitud u homología en un formato relativamente conciso.

1. Acceder a la página principal de *NCBI* (https://www.NCBI.nlm.nih.gov/), dar clic en *BLAST*.

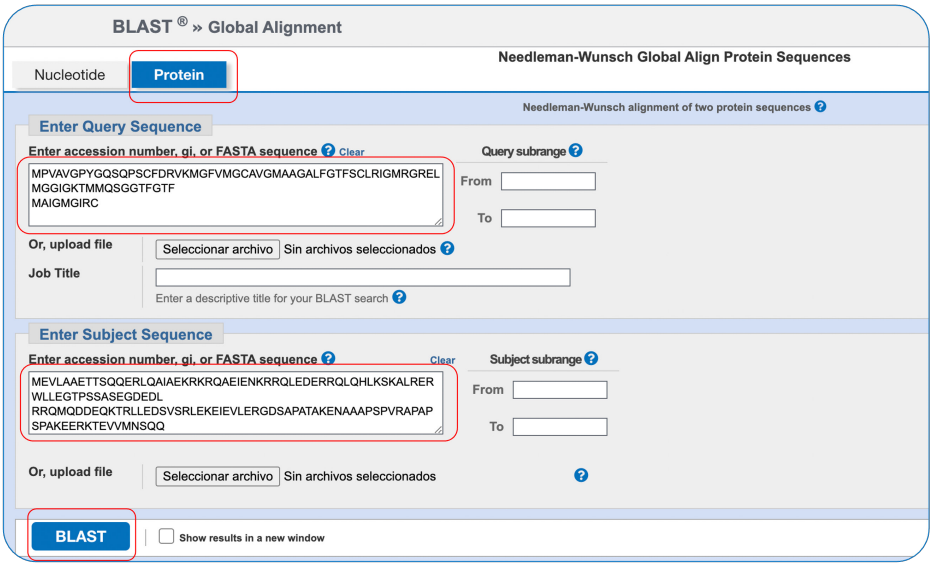


2. Acceder a la opción Alineamiento global, dando clic en "*Global Align*" dentro de la sección "*Specialized Searches*".

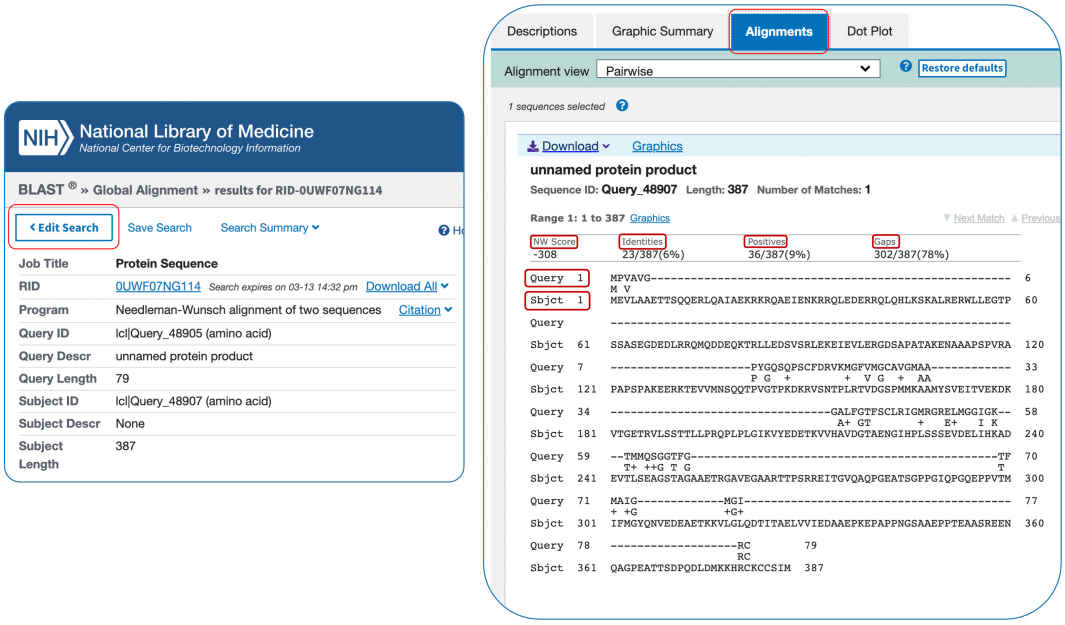


En esta sección se utilizan las secuencias en formato *FASTA* de las dos secuencias relacionadas: >NP_542786.1 *reactive oxygen species modulator* 1 [*Homo sapiens*] y *paralemmin*-1 *isoform* 1 [*Mus musculus*].

Para realizar el alineamiento se introducen las dos secuencias proteicas, también se puede agregar el identificador ("*Accesion*") de cada proteína, sin embargo, se debe quitar el iniciador ">". Dar clic el botón "*Align*".



1. Se muestran los resultados de la siguiente manera:



Datos del recuadro:

- *Query* y *Sbjct*: Se refiere a las dos secuencias, ambas se alinean para maximizar la similitud.
- Coincidencias: Línea entre *Query* y *Sbjct*
- Letra: Indica una coincidencia idéntica
- +: Indica una coincidencia conservada en función de las características químicas del aminoácido.
- Espacio en blanco: No hay coincidencia
- *Score*: Valor del algoritmo de alineamiento
- Identities/Positives: Porcentaje de coincidencias idénticas/conservadas
- Gaps: Porcentaje de huecos incluidos en el alineamiento.

Cuestionario y/o ejercicios complementarios

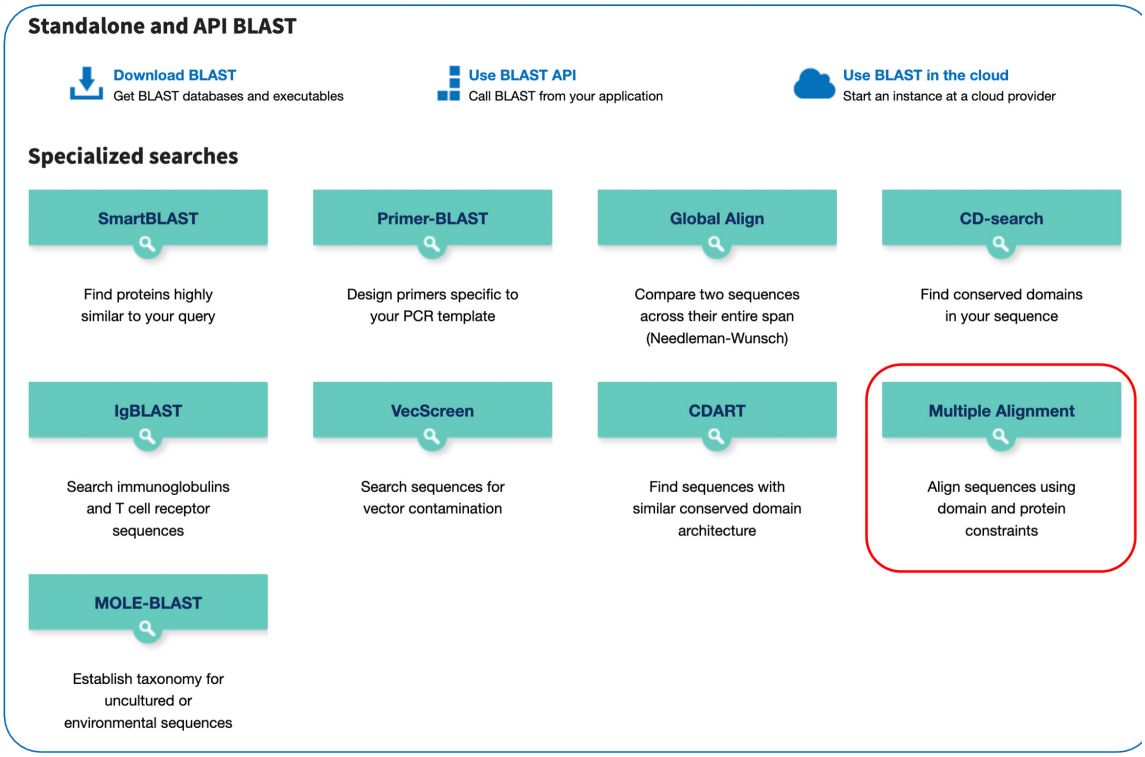
1. Determinar el grado de similitud que hay entre las secuencias del ejercicio.
2. Establecer si existe algún tipo de relación entre ellas (homología y similitud).
3. Identificar los gaps en los alineamientos de las secuencias y relacionarlos con una posible relación evolutiva en la base de datos.

III. Alineamiento de secuencias múltiples

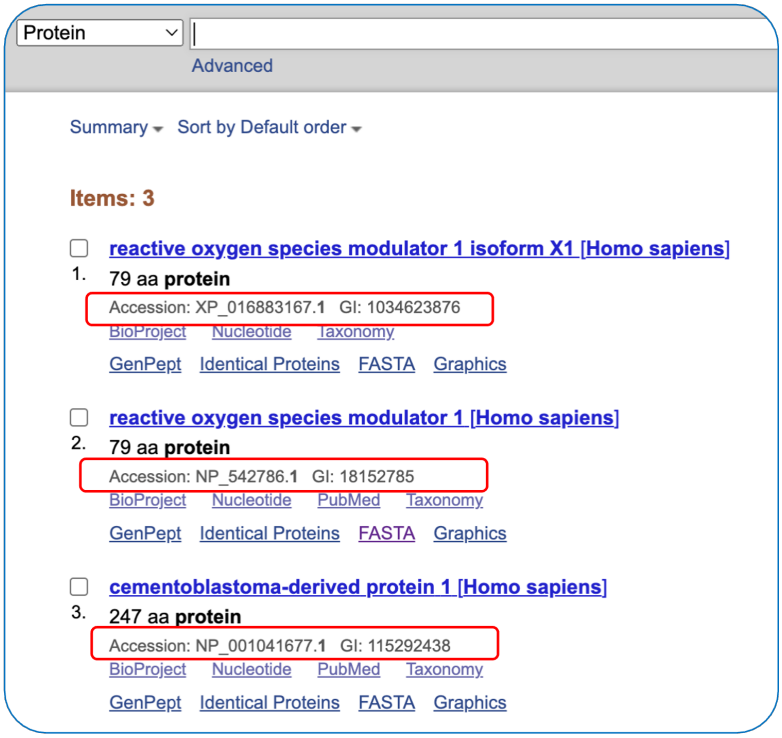
Un alineamiento múltiple de secuencias es aquel que se lleva a cabo con más de dos secuencias de proteínas. El alineamiento múltiple es una de las técnicas bioinformáticas más usadas, ya que por medio de ellas podemos realizar diversos análisis de filogenia, búsqueda de motivos y/o dominios funcionales.

Procedimiento

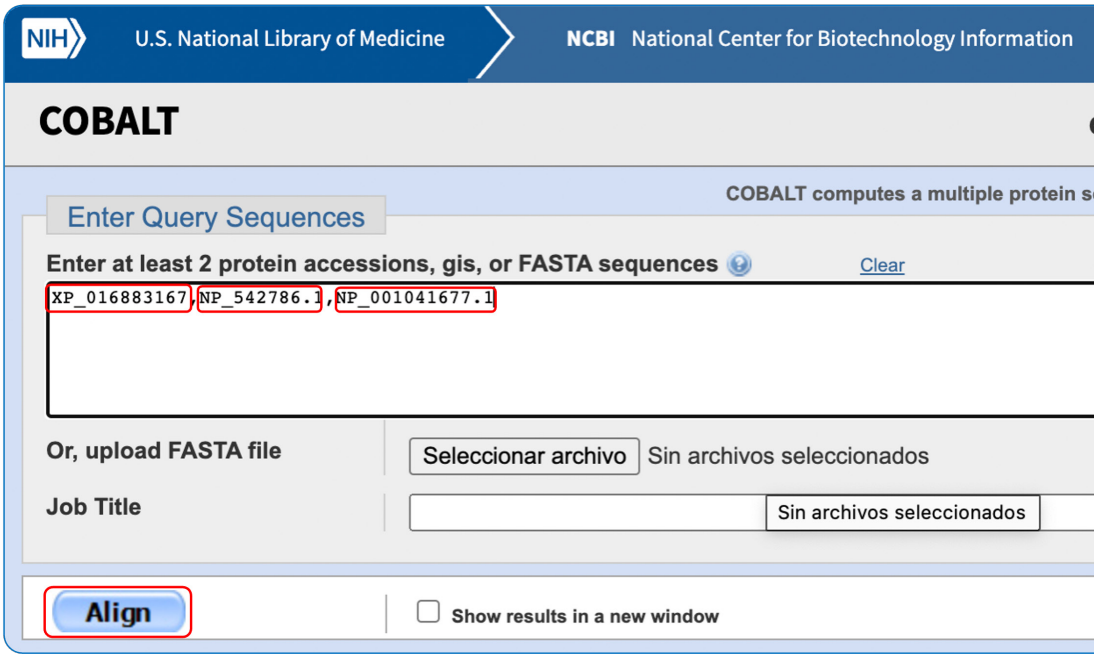
1. Acceder a la página principal de *NCBI* (https://www.NCBI.nlm.nih.gov/), dar clic en *BLAST*, entrar a la opción "*Multiple Alignment*" dentro de la sección "*Specialized Searches*". Se alinearán las secuencias de reactive *oxygen species modulator* 1 *isoform* X1 (*Homo sapiens*), reactive *oxygen species modulator* 1 *isoform* (*Homo sapiens*), *cementoblastoma-derived protein* 1 (*Homo sapiens*).



2. Obtener el identificador (“*Accesion*”) de cada una de las secuencias proteicas a comparar.

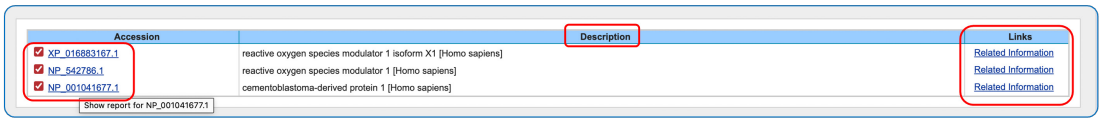


3. Ingresar a la ventana el identificador de cada secuencia seguido de una coma y un espacio y dar clic en “*Align*”.

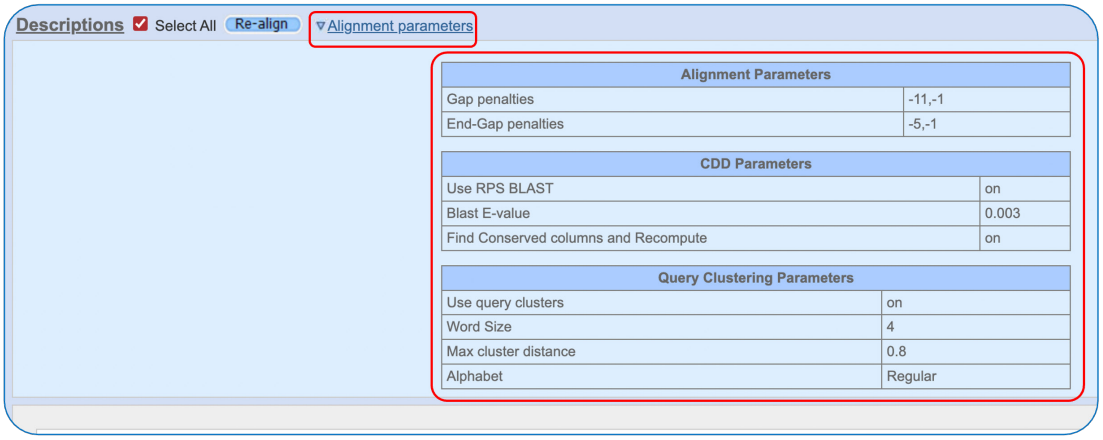


4. De los resultados obtenidos por el alineamiento, dar clic sobre cada una de las representaciones gráficas de las proteínas alineadas.

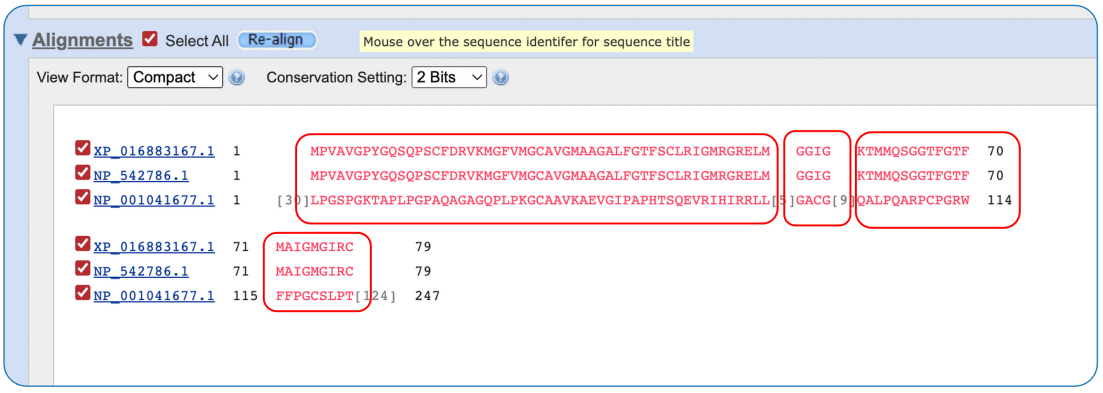
5. Se despliega una descripción general que muestra el número identificador y enlaces en donde podremos obtener información relacionada de cada proteína.



6. Posteriormente, dar clic en la sección parámetros de alineamiento (“*Aligment parameters*”) para obtener el *E-value*.



7. En la sección de alineamiento se visualizan las tres secuencias de proteínas alineadas donde se observan los *match* y *gaps* específicos. Puede cambiar el formato de alineamiento en compacto, medio y expandido, que se encuentra más detallado.



Cuestionarioy/o ejercicios complementraios

1. Determinar el grado de similitud que hay entre las secuencias del ejercicio.
2. Identificar si existe algún tipo de relación.
3. ¿Qué indica el E-value?
4. ¿Por qué comparar secuencias y no la estructura de las proteínas?
5. ¿Por qué no se usa la función de las proteínas para predecir la homología?

Bibliografía

- Atwood, T., y Parry-Smith, D. (2002). *Introducción a la bioinformática*. Madrid, España: Pearson Educación, S.A.
- Capel, J., y Yuste, F. (2016). *Manual de prácticas de bioinformática*. Andalucía, España: Almería.
- Lodish, H., Berk, A., Kaiser, C., Kriegeer, M., Bretscher, A., Ploegh, H., Amon, A., y Scott, M. (2016). *Biología celular y molecular.* Buenos Aires, Argentina: Médica Panamericana.
- Oliva, R., y Vidal, J. (2006). *El genoma humano: nuevos avances en investigación, diagnóstico y tratamiento*. Barcelona, España: UBe.
- Roldan, D. (2015). *Bioinformática, el ADN en un solo clic*. Bogotá, Colombia: Ediciones de la U. Stryer, L., Berg, J., y Tymokzko, J. (2013). Bioquímica. Madrid, España: Reverté, S.A.
- Werner, M. (2008). *Bioquímica. Fundamentos para medicina y ciencias de la vida*. Barcelona, España: Reverté.

Práctica 3

Comparación de secuencias de aminoácidos

T-COFFEE

T-COFFEE

T-Coffee (*Tree-based Consistency Objective Function for alignment Evaluation*) es un algoritmo progresivo de alineamiento múltiple de secuencias basado en consistencia que realiza una matriz de distancias (valores de similitud de secuencias) y un árbol guía para ver la relación entre secuencias. Una vez construido el árbol guía, las secuencias se alinean de forma progresiva siguiendo las ramas del árbol y uniéndose en grupos de dos secuencias al inicio y alineamientos intermedios posteriormente.

Objetivo general

- Comparar dos o más secuencias de proteínas mediante el programa *T-Coffee* con el fin de detectar similitudes.

Objetivos particulares

1. Contar con elementos teóricos para analizar y diferenciar la identidad, homología o similitud de proteínas y predecir la estructura y función.

2. Detectar la homología o similitud entre secuencias de las proteínas mediante su alineamiento para predecir la función.

Requerimientos para la elaboración de la práctica

- Computadora con conexión a Internet.

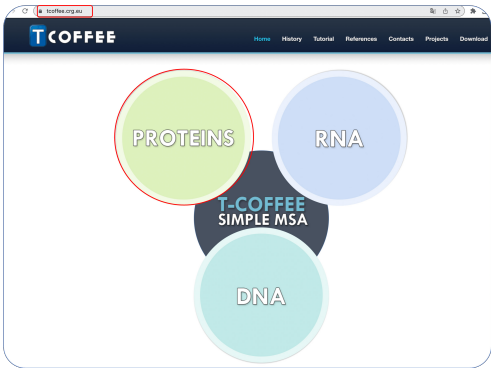
Nota: No es necesario descargar ningún programa para realizar esta práctica, se trabaja en línea.

Procedimientos

I. Obtención del formato *FASTA* de una proteína en *NCBI*
Obtener el formato *FASTA*, realizar el procedimiento de la primera sección de la práctica 2.

II. Alineamiento de dos o más secuencias con *T-Coffee*
El poder del análisis de secuencias reside en la capacidad de tomar simultáneamente secuencias relacionadas y expresar el grado de similitud u homología en un formato relativamente conciso.

1. Acceder al programa *T-Coffee* en la siguiente liga http://tcoffee.crg.cat/, dar clic en "*Proteins*".



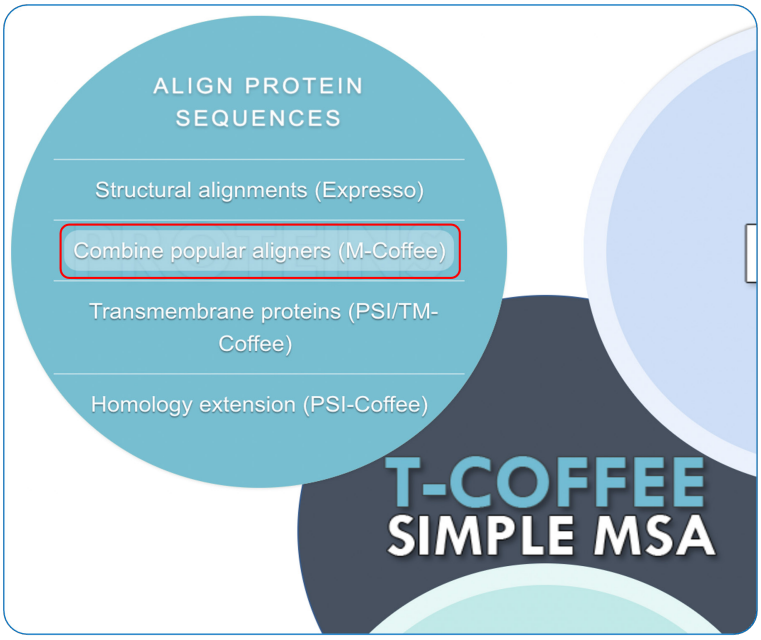
2. Posteriormente, se desplegará un panel de opciones, las cuales son:

   - *Structural alignments* (*Expresso*): Alinea secuencias de proteínas usando información estructural que se encuentra libre en Internet, es precisa sobre todo cuando las proteínas a comparar tienen una estructura en 3 dimensiones conocida.

   - *Combine popular aligners* (*M-Coffee*): Esta opción combina los resultados de varios métodos de alineamiento mostrando las porciones en las cuales estos coinciden.

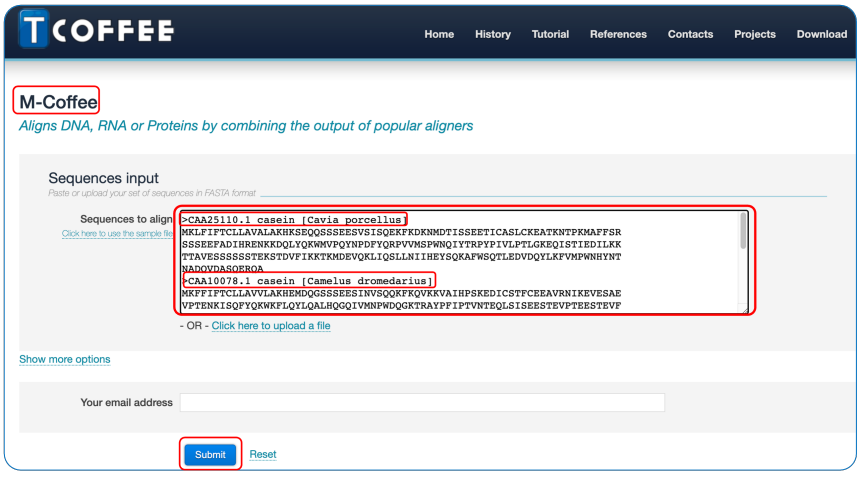
   - *PSI-Coffee*: Realiza el alineamiento determinando homologías, es más lento para alinear proteínas que podrían estar relacionadas. También se puede ocupar este método de alineamiento específicamente para proteínas transmembranales

3. En esta práctica se utilizará el alineamiento por varios métodos, dando clic en "*Combine popular aligners* (*M-Coffee*)".

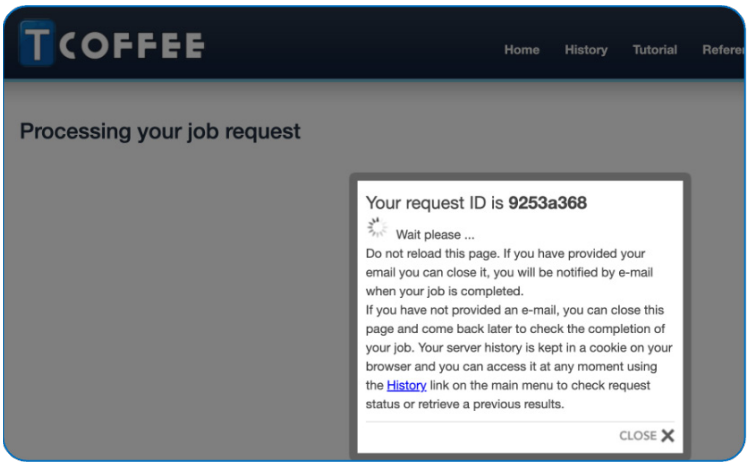


En esta sección se trabajará con las secuencias en formato *FASTA* de 4 secuencias relacionadas. Para realizar el ejercicio se usarán: >CAA25110.1 *casein* [*Cavia porcellus*], >CAA10078.1 *casein* [*Camelus dromedarius*], >AAY68392.1 *casein* [*Homo sapiens*], >AEN87275.1 *casein alpha/beta* [*Bacillus megaterium* WSH-002].

4. Para realizar el alineamiento se introducen las secuencias proteicas en formato *FASTA*, incluir el iniciador ">" y el identificador ("*Accesion*") de cada proteína. Dar clic el botón "*Submit*".

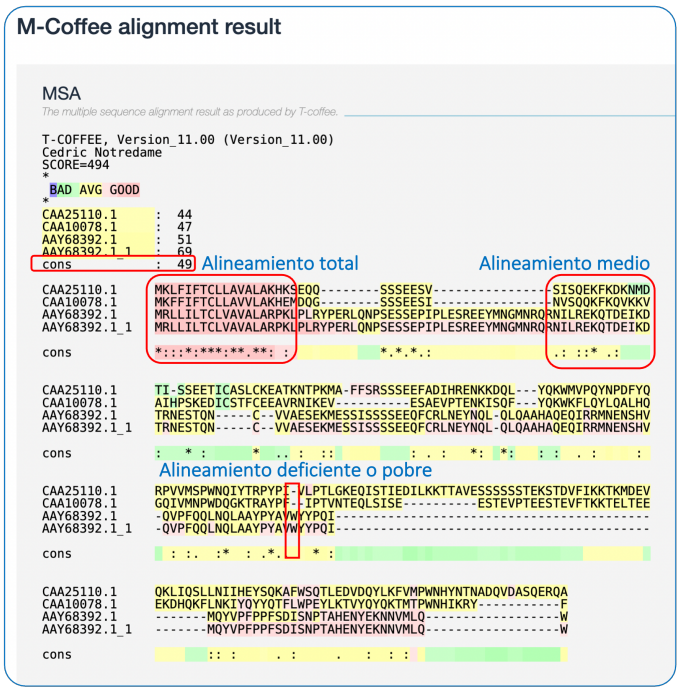


5. Los resultados pueden tardar unos minutos, durante los cuales no se debe actualizar la página.



6. Los resultados obtenidos muestran un alineamiento que contiene códigos de color de acuerdo con las coincidencias del alineamiento realizado por varios métodos, donde:

   - Las regiones marcadas en rojo corresponden a un alineamiento total por todos los métodos.
   - Las regiones en azul corresponden a un alineamiento pobre.
   - Las regiones en verde y en amarillo deben analizarse con precaución, sobre todo si las secuencias serán utilizadas para analizar filogenia.
   - En la parte superior del análisis hay una sección que indica la consistencia "*cons*" por secuencia, yendo de 0 a 100, la cual indica la confiabilidad del alineamiento. Una medida menor a 50 indica una coincidencia pobre en el alineamiento.
   - Espacios punteados: no hay coincidencia
   - *Gaps*: porcentaje de huecos incluidos en el alineamiento.



En la parte inferior de los resultados se muestra una sección en la cual podemos descargar los alineamientos realizados por cada método o programa que utiliza *T-Coffee*, así como el árbol guía para realizar la comparación de secuencias.

Cuestionario y/o ejercicios complementarios

1. ¿Cuáles son las ventajas de utilizar *T-Coffee* en lugar de otros programas?
2. Investigar qué indican los *gaps* en los alineamientos de las secuencias (funcional, estructural y evolutivamente).
3. ¿A qué se debe la coincidencia pobre en el alineamiento?
4. Realiza el alineamiento de secuencias de las siguientes proteínas:

interferon-induced transmembrane protein 2 [*Rattus norvegicus*], *interferon-induced transmembrane protein* 3 [*Homo sapiens*], *interferon-induced transmembrane protein* 2 [*Homo sapiens*], *interferon-induced transmembrane protein* 1 [*Homo sapiens*].

¿Qué opción de alineamiento utilizaría para la comparación de las secuencias?, ¿por qué? y ¿qué valor de consistencia resulta?

Bibliografía

- Atwood, T., y Parry-Smith, D. (2002). *Introducción a la bioinformática*. Madrid, España: Pearson Educación, S.A.
- Capel, J., y Yuste, F. (2016). *Manual de prácticas de bioinformática*. Andalucía, España: Almería.
- Lodish, H., Berk, A., Kaiser, C., Kriegeer, M., Bretscher, A., Ploegh, H., Amon, A., y Scott, M. (2016). *Biología celular y molecular*. Buenos Aires, Argentina: Médica Panamericana.
- Notredame, C., Higgins, G., Heringa, J. (2000). "*T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment*". J. Mol. Biol. 302: 205-17.
- Roldan, D. (2015). *Bioinformática, el ADN en un solo clic*. Bogotá, Colombia: RA-MA ediciones de la U.

Práctica 4

Del gen a la proteína

ExPASy

Introducción

El flujo de información genética se da a nivel celular, el DNA del genoma no dirige directamente la síntesis de proteínas, sino que utiliza el RNA como intermediario. Cuando la célula necesita una proteína determinada, la secuencia de nucleótidos de la región apropiada de la inmensamente larga molécula de DNA se copia primero a RNA, que constituye el primer paso del flujo de información conocido como *transcripción* el cual permite obtener a partir de un transcrito primario de RNAm o pre-RNAm distintas moléculas de RNAm maduras. En el segundo paso del flujo de información, estas copias de RNAm se utilizan como moldes para dirigir la síntesis de las proteínas, proceso conocido como traducción. Por lo tanto, el flujo de información genética en las células va de DNA a RNA y de éste a proteínas.

Todas las células expresan la información genética de esta manera -un principio fundamental que se ha denominado el *dogma central* de la biología molecular-. A pesar de la universalidad de este dogma, existen importantes variaciones en el mismo por el que la información fluye desde el DNA a las proteínas. Una de las principales es que, en las células eucariotas, los transcritos de RNA sufren una serie de procesamientos en el núcleo (incluyendo su maduración) antes de salir al citoplasma y ser traducido a proteínas. En esta etapa del procesamiento pueden cambiar aspectos críticos del "significado" de la molécula de RNA y por tanto resulta crucial para entender de qué manera las células eucariotas leen sus genomas.

La síntesis de todas las cadenas polipeptídicas de una célula eucariota o procariota comienza con el aminoácido metionina. En las bacterias se emplea una forma especializada de la metionina con un grupo formilo que se encuentra unido a un grupo amino. En la mayoría de los RNAm, el codón de inicio es el AUG, en unos pocos RNAm bacterianos, el codón de inicio GUG y ocasionalmente se usa como codón de inicio para metionina en los eucariotas CUG. Los tres codones UAA, UGA y UAG no especifican aminoácidos, sino que constituyen codones de terminación que marcan el extremo carboxilo de las cadenas polipeptídicas de la mayoría de las células. La secuencia de codones que va entre un codón de inicio específico hasta un codón de término se llama marco de lectura.

Esta disposición lineal precisa de ribonucleótidos en grupo de tres, en el RNAm, especifica la secuencia lineal precisa de aminoácidos en una cadena polipeptídica, y también señala donde se inicia y termina la síntesis de la cadena. Dado que el código genético, es un código de tripletes que no se superpone, sin divisiones entre ellos, un RNAm particular, en teoría debería traducirse en tres marcos de lectura en cada dirección 5´y 3´. A la región que codifica a una proteína se le llama región codificante del gen (*CDS* por sus siglas en inglés) o "marco de lectura abierto".

ExPASy

ExPASy es el portal de recursos bioinformáticos del Instituto Suizo de Bioinformática (*SIB*), que proporciona acceso a bases de datos científicas y herramientas informáticas en diferentes áreas de la ciencia, incluyendo proteómica, genómica, filogenia, sistemas biológicos, genética de poblaciones, transcriptómica y otros. También nos da acceso a otros recursos del SIB e instituciones externas.

Para realizar esta práctica se utilizará la herramienta "*Translate*" de *ExPASy*, que nos permite obtener la secuencia de aminoácidos a partir de la secuencia codificante del DNA.

Objetivos generales

- Mediante herramientas bioinformáticas identificar el flujo de información de DNA a RNA para determinar regiones codificantes capaces de sintetizar un polipéptido o una proteína.

Objetivos particulares

1. Identificar una secuencia de DNA codificante en una base de datos y utilizarla como molde para generar el RNA maduro y su posterior traducción a proteína.

2. Analizar las 6 secciones posibles de la traducción a aminoácidos con el fin de determinar el fragmento que sea funcional.

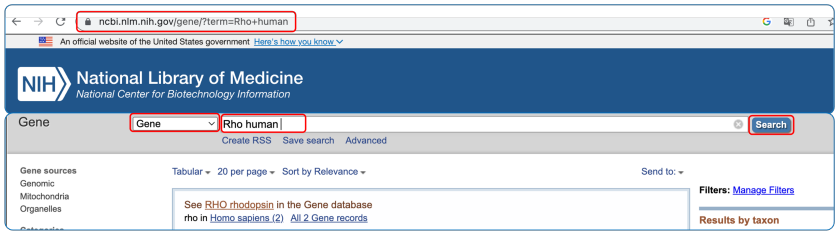
Requerimientos para la práctica

- Dispositivo con conexión a Internet.

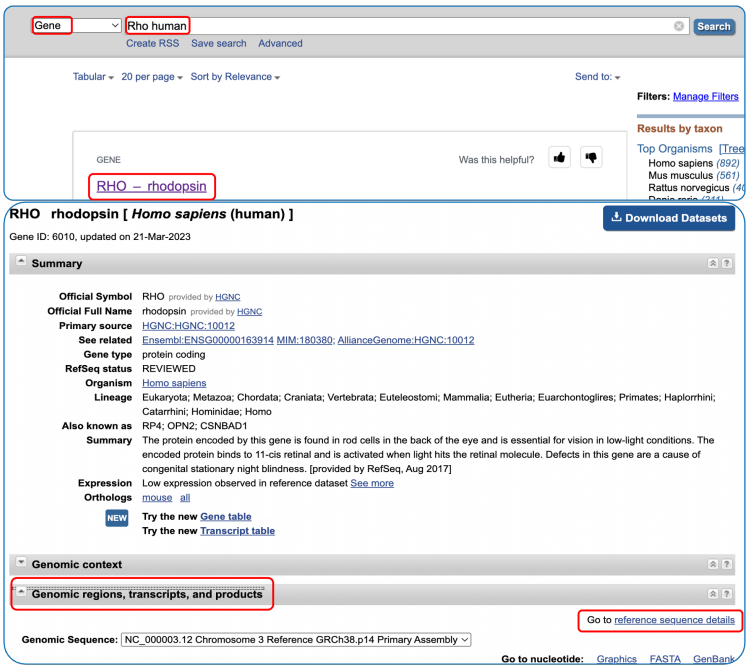
Nota: No es necesario descargar ningún programa para realizar esta práctica, se trabaja en línea.

Procedimiento

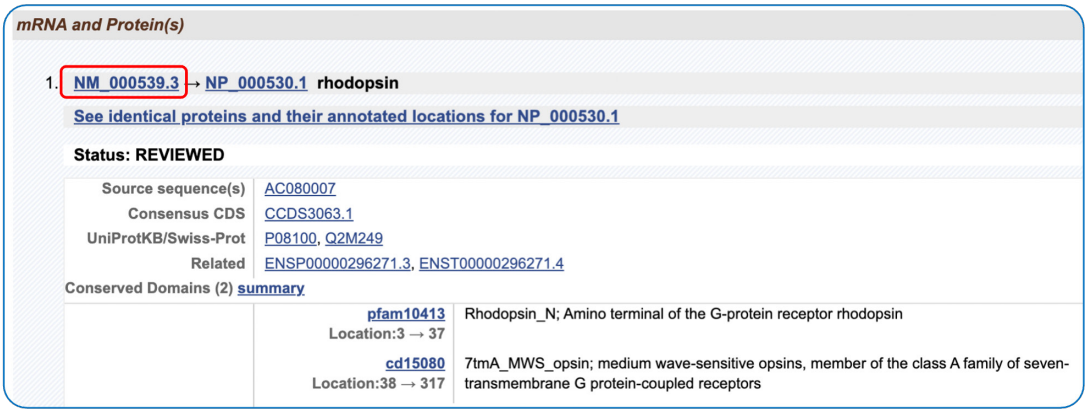
1. Acceder a NCBI (https://www.NCBI.nlm.nih.gov/), en el botón "*All Databases*" y elegir la opción "*Gene*", ingresar a la ventana el nombre de la proteína ejemplo RHO *rhodopsin* [*Homo sapiens* (*human*)]. Dar clic en "*Search*".



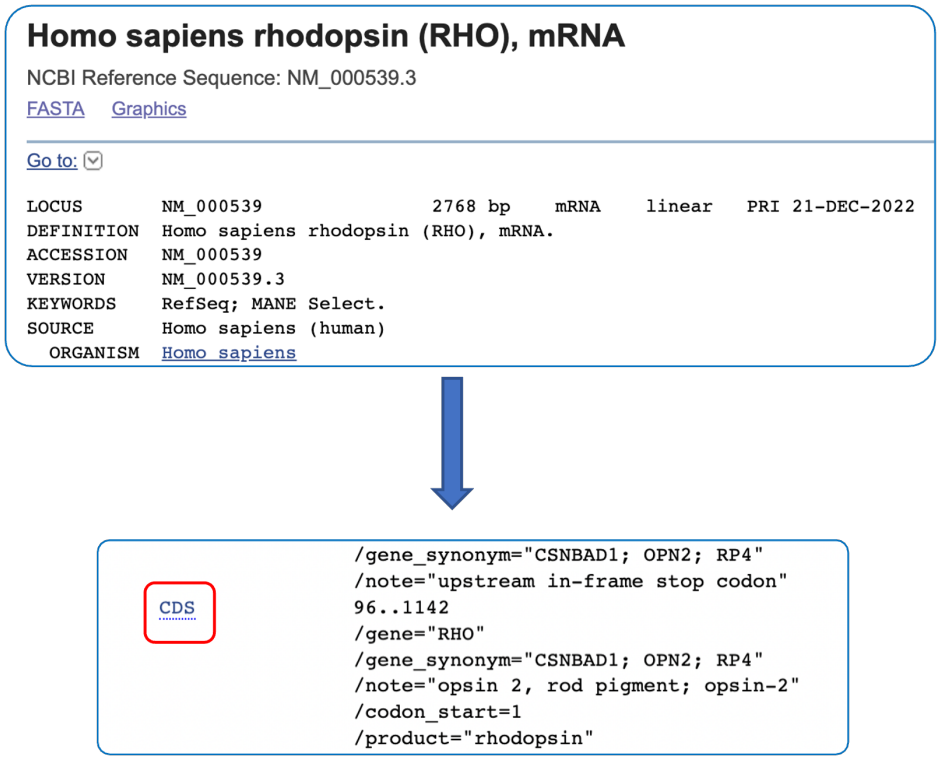
2. Se mostrará la información del gen, dar clic a la sección de los detalles de las regiones genómicas, transcritos y productos ("*Genomic regions/Reference sequence details*").



3. Dar clic en la sección correspondiente al RNAm (en esta sección se tienen los datos experimentales de los procesos de corte y empalme o *splicing* alternativo que han sido reportadas para este gen).



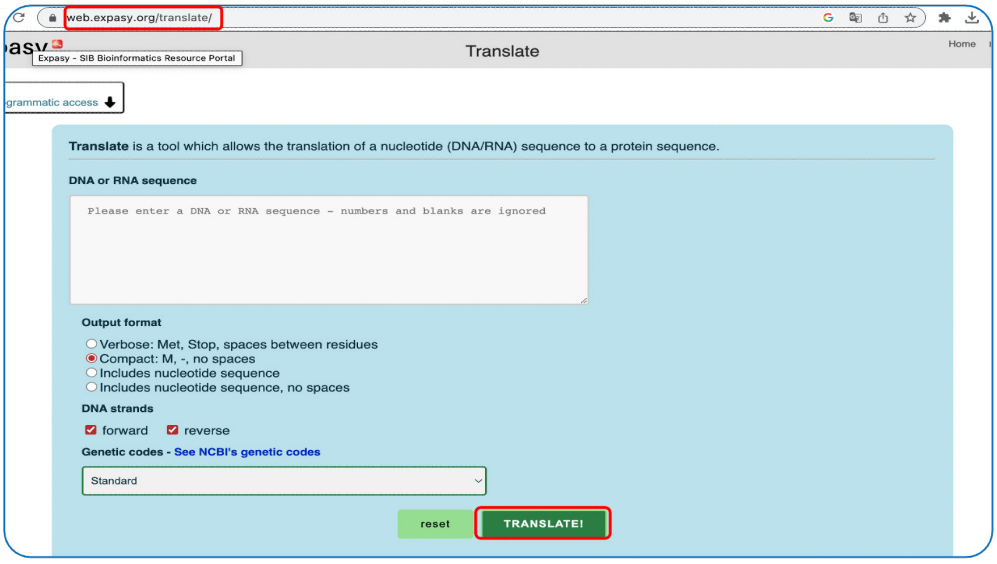
4. Se desplegará información diversa, en la cual se busca la región codificante o *CDS* ("*Coding Sequence*"), dar clic.



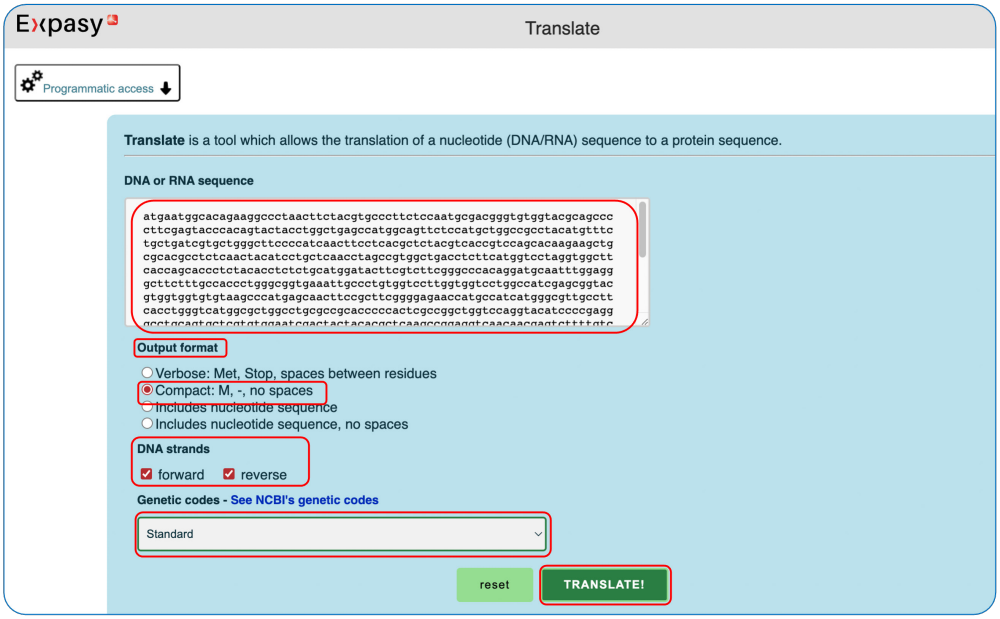
5. La región codificante de la secuencia del gen se muestra marcada, copiar esta secuencia.

```
ORIGIN
        1 agagtcatcc agctggagcc ctgagtggct gagctcaggc cttcgcagca ttcttgggtg
       61 ggagcagcca cgggtcagcc acaagggcca cagccatgaa tggcacagaa ggccctaact
      121 tctacgtgcc cttctccaat gcgacgggtg tggtacgcag ccccttcgag tacccacagt
      181 actacctggc tgagccatgg cagttctcca tgctggccgc ctacatgttt ctgctgatcg
      241 tgctgggctt ccccatcaac ttcctcacgc tctacgtcac cgtccagcac aagaagctgc
      301 gcacgcctct caactacatc ctgctcaacc tagccgtggc tgacctcttc atggtcctag
      361 gtggcttcac cagcaccctc tacacctctc tgcatggata cttcgtcttc gggcccacag
      421 gatgcaattt ggagggcttc tttgccaccc tgggcggtga aattgccctg tggtccttgg
      481 tggtcctggc catcgagcgg tacgtggtgg tgtgtaagcc catgagcaac ttccgcttcg
      541 gggagaacca tgccatcatg ggcgttgcct tcacctgggt catggcgctg gcctgcgccg
      601 caccccact cgccggctgg tccaggtaca tccccgaggg cctgcagtgc tcgtgtggaa
      661 tcgactacta cacgctcaag ccggaggtca acaacgagtc ttttgtcatc tacatgttcg
      721 tggtccactt caccatcccc atgattatca tctttttctg ctatgggcag ctcgtcttca
      781 ccgtcaagga ggccgctgcc cagcagcagg agtcagccac cacacagaag gcagagaagg
      841 aggtcacccg catggtcatc atcatggtca tcgctttcct gatctgctgg gtgccctacg
      901 ccagcgtggc attctacatc ttcacccacc agggctccaa cttcggtccc atcttcatga
      961 ccatcccagc gttctttgcc aagagcgccg ccatctacaa ccctgtcatc tatatcatga
     1021 tgaacaagca gttccggaac tgcatgctca ccaccatctg ctgcggcaag aacccactgg
     1081 gtgacgatga ggcctctgct accgtgtcca agacggagac gagccaggtg gccccggcct
     1141 aagacctgcc taggactctg tggccgacta taggcgtctc ccatccccta caccttcccc
     1201 cagccacagc catcccacca ggagcagcgc ctgtgcagaa tgaacgaagt cacataggct
     1261 ccttaatttt ttttttttt ttaagaaata attaatgagg ctcctcactc acctgggaca
     1321 gcctgagaag ggacatccac caagacctac tgatctggag tcccacgttc cccaaggcca
     1381 gcgggatgtg tgcccctcct cctcccaact catctttcag gaacacgagg attcttgctt
     1441 tctggaaaag tgtcccagct tagggataag tgtctagcac agaatggggc acacagtagg
     1501 tgcttaataa atgctggatg gatgcaggaa ggaatggagg aatgaatggg aagggagaac
     1561 atatctatcc tctcagaccc tcgcagcagc agcaactcat acttggctaa tgatatggag
     1621 cagttgtttt tccctccctg ggcctcactt tcttctccta taaaatggaa atcccagatc
     1681 cctggtcctg ccgacacgca gctactgaga agaccaaaag aggtgtgtgt gtgtctatgt
     1741 gtgtgtttca gcactttgta aatagcaaga agctgtacag attctagtta atgttgtgaa
     1801 taacatcaat taatgtaact agttaattac tatgattatc acctcctgat agtgaacatt
     1861 ttgagattgg gcattcagat gatggggttt cacccaacct tggggcaggt ttttaaaaat
     1921 tagctaggca tcaaggccag accagggctg ggggttgggc tgtaggcagg gacagtcaca
     1981 ggaatgcaga atgcagtcat cagacctgaa aaaacaacac tgggggaggg ggacggtgaa
     2041 ggccaagttc ccaatgaggg tgagattggg cctggggtct caccctagt gtggggcccc
     2101 aggtcccgtg cctcccctc ccaatgtggc ctatggagag acaggccttt ctctcagcct
     2161 ctggaagcca cctgctcttt tgctctagca cctgggtccc agcatctaga gcatggagcc
     2221 tctagaagcc atgctcaccc gcccacattt aattaacagc tgagtccctg atgtcatcct
     2281 tatctcgaag agcttagaaa caaagagtgg gaaattccac tgggcctacc ttccttgggg
     2341 atgttcatgg gccccagttt ccagtttccc ttgccagaca agcccatctt cagcagttgc
     2401 tagtccattc tccattctgg agaatctgct ccaaaaagct ggccacatct ctgaggtgtc
     2461 agaattaagc tgcctcagta actgctcccc cttctccata taagcaaagc cagaagctct
     2521 agctttaccc agctctgcct ggagactaag gcaaattggg ccattaaaag ctcagctcct
     2581 atgttggtat taacggtggt gggttttgtt gctttcacac tctatccaca ggatagattg
     2641 aaactgccag cttccacctg atccctgacc ctgggatggc tggattgagc aatgagcaga
     2701 gccaagcagc acagagtccc ctggggctag aggtggagga ggcagtcctg ggaatgggaa
     2761 aaacccca
```

6. Acceder a ExPAsy https://web.expasy.org/translate/. Dar clic en "*Translate*".

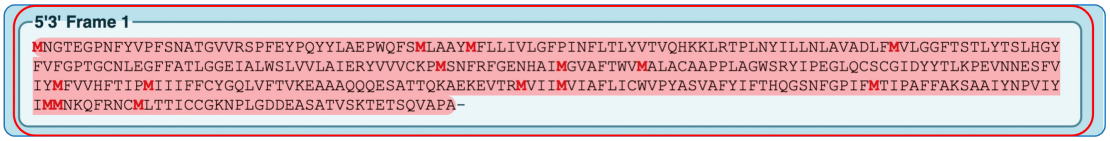


7. Ingresar la *CDS* del gen en la ventana. Se eligen los siguientes parámetros: formato de salida compacto ("*Output format*"), cadenas de DNA *Forward* y *Reverse* ("DNA *strands*") y código genético estándar ("*Genetic codes*"). Dar clic en "*Translate*".



8. En la secuencia traducida se muestra el codón de inicio metionina (**M**) y los codones de paro (*stop*) en las 6 secciones, como resultado de la lectura en ambos sentidos de la cadena de DNA (*upstream*: contracorriente o río arriba 5´a 3´, o *downstream*: río abajo o 3´a 5´).

9. En este caso, la región mejor traducida es el fragmento más largo y resaltado, que corresponde a la secuencia 5′ a 3′ de la primera sección.



Cuestionario y/o ejercicios complementarios

1. Describir el concepto de dogma central de la biología molecular tomando como base la práctica.
2. ¿Es correcto llamarle "dogma"?
3. ¿A qué se refiere el *CDS* y por qué no ocupar la secuencia del gen completo?
4. Investigar que es un marco abierto de lectura (*open reading frame*).
5. ¿Por qué resultan 6 secciones de la traducción a aminoácidos en *ExPASy*?

Bibliografía

- Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., y Walter, P. (2016). *Biología molecular de la célula. Barcelona*, España: Omega.
- Atwood, T., y Parry-Smith, D. (2002). *Introducción a la bioinformática*. Madrid, España: Pearson educación, S.A.
- Gasteiger E., Gattiker A., Hoogland C., Ivanyi I., Appel R.D., Bairoch A. (2003). *ExPASy: the proteomics server for indepth protein knowledge and análisis*. Nucleic Acids Res. 31:3784-3788
- Gómez, J., y Gómez, C. (2004). *Iniciación al estudio de la Bioquímica*. Madrid, España: Grupo Anaya, S.A.
- Macarulla, J., y Goñi, F. (2013). *Biomoléculas*. Barcelona, España: Reverté, S.A.
- Melo, V., y Cuamatzi, O. (2004). *Bioquímica de los procesos metabólicos*. Barcelona, España: Reverté, S.A.
- Oliva, R., y Vidal, J. (2006). *El genoma humano: nuevos avances en investigación, diagnóstico y tratamiento*. Barcelona, España: UBe.
- Peña, A., Arroyo, A., Gómez, A., y Tapia, R. (2018). *Bioquímica*. Ciudad de México, México: Limusa.
- Roldan, D. (2015). *Bioinformática, el ADN en un solo clic*. Bogotá, Colombia: Ediciones de la U.

Práctica 5

Predicción de las propiedades fisioquímicas

PROTPARAM/UNIPROT

Introducción

Las propiedades fisicoquímicas de las proteínas están determinadas por las características intrínsecas de los aminoácidos, por tal motivo las proteínas poseen diferentes propiedades que ejercen impactos críticos sobre su actividad, estructura y función biológica. Estas propiedades se pueden calcular y predecir para comprender mejor la función o la actividad de una proteína.

Existen proteínas solubles en agua o disoluciones salinas; otras en cambio, requieren márgenes críticos de concentración salina; hay proteínas solubles sólo en disoluciones básicas, ácidas o hidroalcohólicas; algunas otras son totalmente insolubles en medio acuoso. Estas diferencias de solubilidad son muy útiles a la hora de purificar proteínas.

La carga total de la molécula proteica depende del pH de la solución y del número relativo de cada aminoácido en la molécula. Así, cuando el pH de una solución es tal que la carga neta de la molécula es 0, es decir cuando el número total de cargas negativas es igual al número total de cargas positivas presentes en la molécula, se denomina punto isoeléctrico (pI) o pH isoeléctrico.

ProtParam

ProtParam es una herramienta que permite el cálculo de varios parámetros fisicoquímicos de una proteína conocida y almacenada en *Swiss-Prot* o *TrEMBL*, o para una secuencia de proteínas ingresada por el usuario. Los parámetros calculados incluyen el peso molecular, pI teórico, composición de aminoácidos, composición atómica, coeficiente de extinción, vida media estimada, índice de estabilidad, índice alifático y promedio de hidrofobicidad.

UniProt

Universal Protein Resource (*UniProt*) proporciona un recurso central, completo y de libre acceso sobre secuencias de proteínas y anotaciones funcionales, es un recurso central para almacenar e interconectar información de fuentes grandes y dispares, y es el catálogo más completo de secuencias de proteínas y anotación funcional.

Objetivo general

- Identificar las propiedades fisicoquímicas de las proteínas mediante la herramienta *ProtParam* de *ExPASy* con el fin de contar con elementos físicos y químicos para el estudio de proteínas de interés.

Objetivos particulares

1. Obtener los parámetros fisicoquímicos estimados de un dominio o fragmentos de proteína mediante *ProtParam* para conocer las características de la secuencia elegida.

2. Analizar las características fisicoquímicas mediante una base de datos para su estudio.

Requerimientos para la práctica

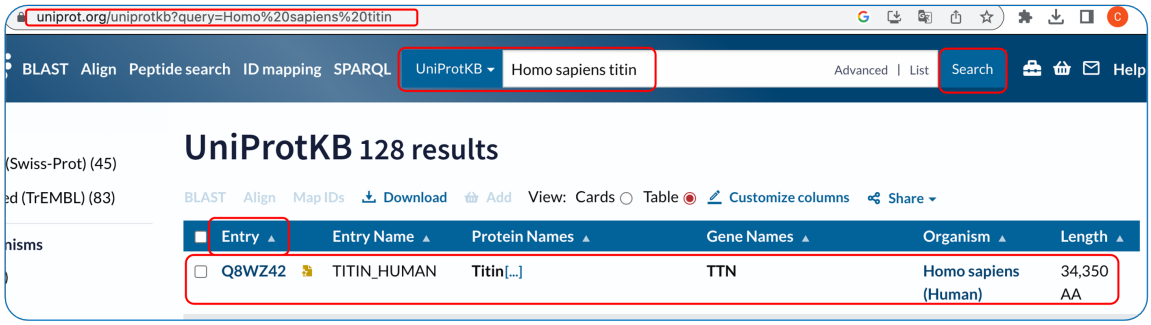
- Dispositivo con conexión a Internet.

Nota: No es necesario descargar ningún programa para realizar esta práctica, se trabaja en línea.

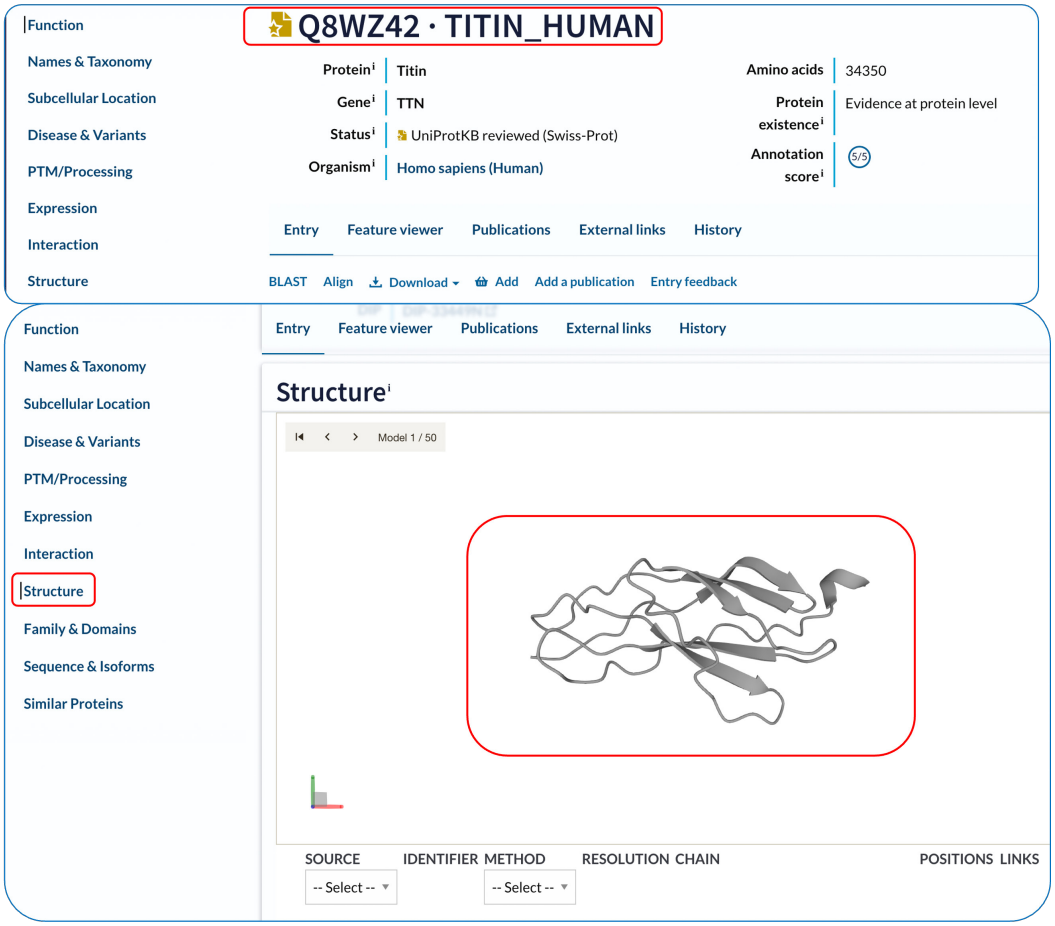
Procedimiento

1. En esta práctica se tomará como ejemplo la proteína titina (*titin* [*Homo sapiens*], *GenBank*: CAA62188.1). Se busca la secuencia en *NCBI* (*NCBI/protein/* (nombre de la proteína) */FASTA*).

2. Buscar la proteína en *Uniprot* (https://www.uniprot.org/) y buscar el número identificador (*Entry: Unique and stable entry identifier*).

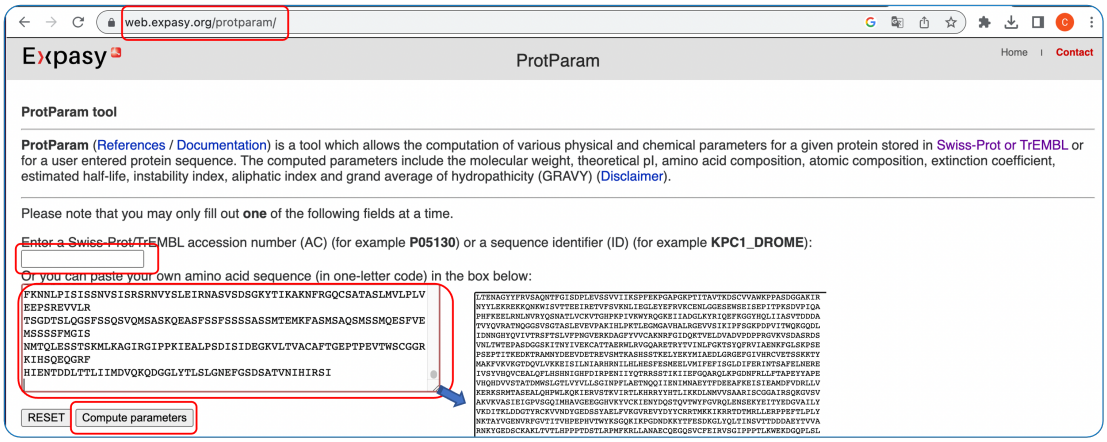


3. Dar clic en el número identificador para obtener la información de la proteína. La base de datos para el análisis se encuentra descrito en la siguiente imagen, del lado izquierdo: la función, nombre, taxonomía, localización subcelular, modificaciones postraduccionales, familia, dominios, secuencia y estructura.

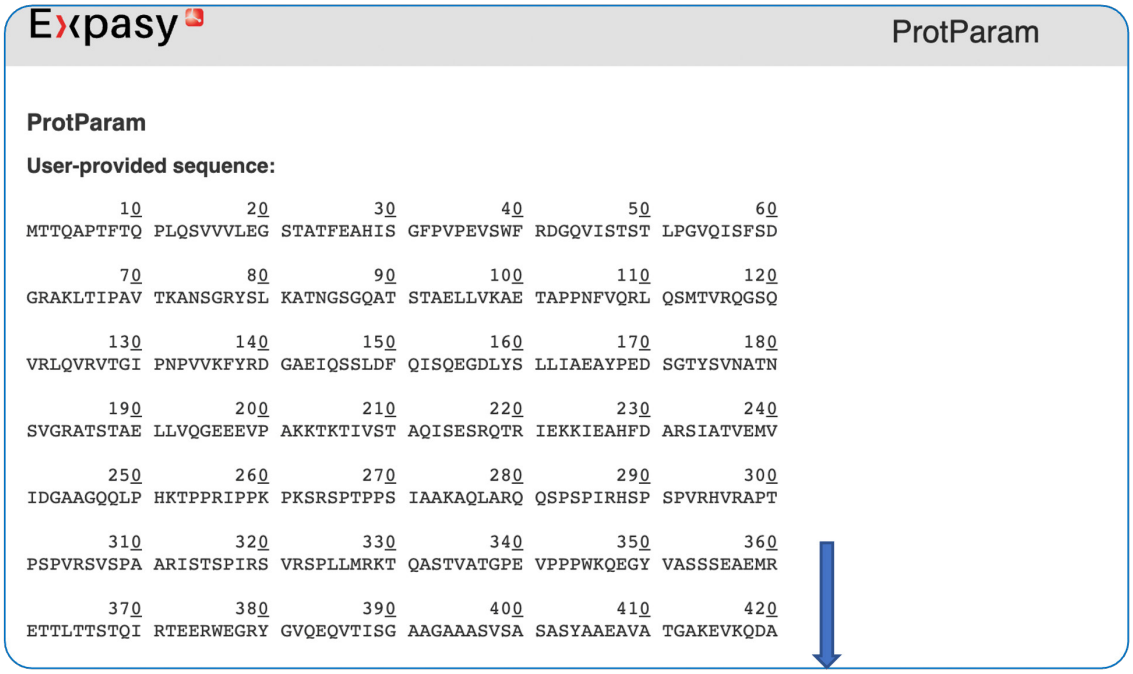


4. Para obtener las características fisicoquímicas se accede a la página http://expasy.org/tools/protparam.html.

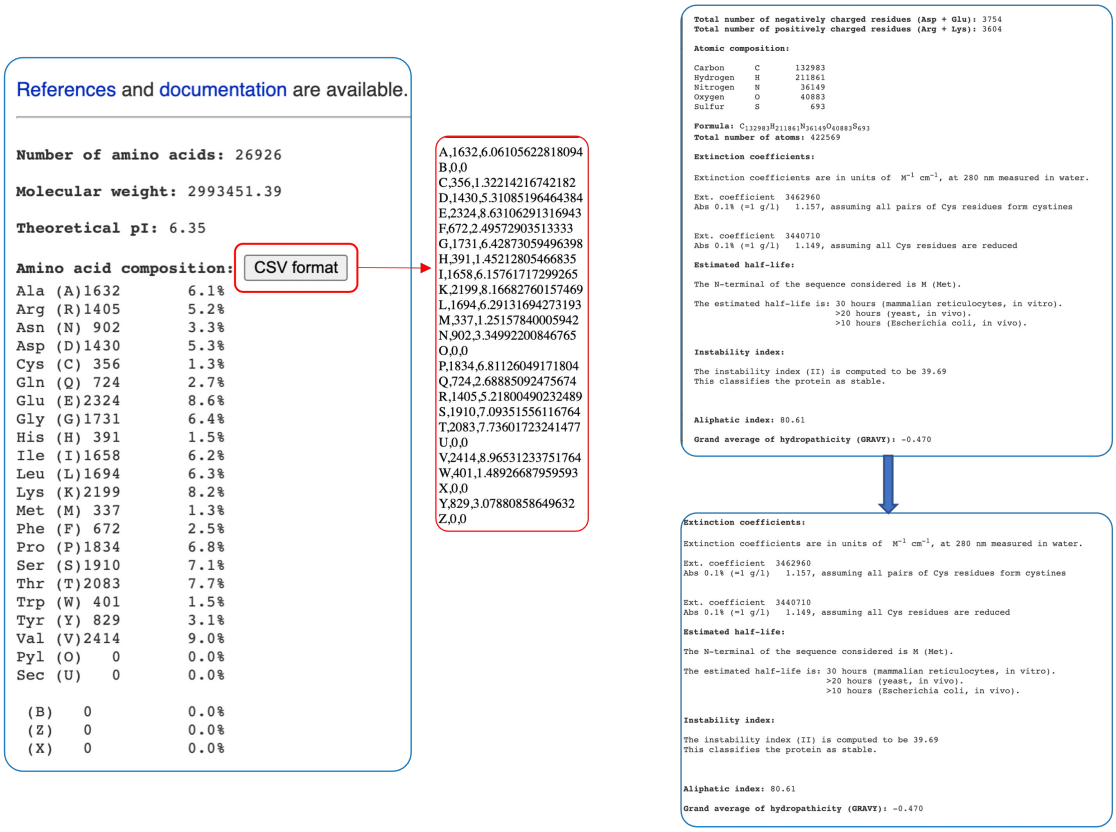
5. Copiar el número de acceso (“*accesion*”) y pegarla en la ventana pequeña, o la secuencia completa de la proteína en la ventana grande. En este ejemplo se utilizó la secuencia completa. Dar clic en “*Compute parameters*”.



6. Como resultado se muestra la secuencia introducida seguido de los parámetros a analizar.



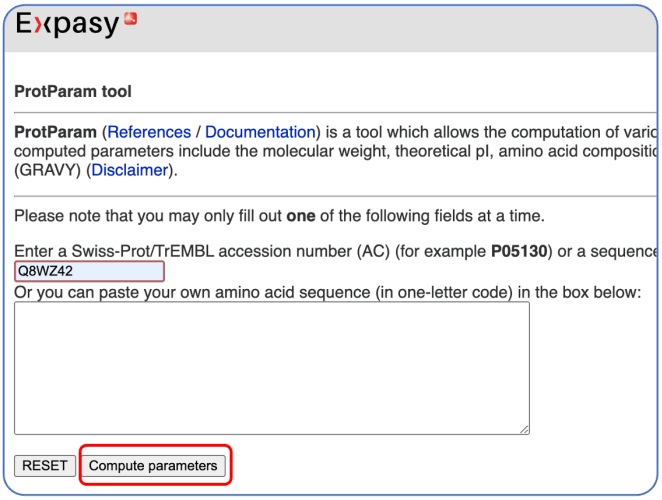
ProtParam de *ExPASy* permite calcular parámetros fisicoquímicos de una secuencia proteica dada, entre los que se encuentran el peso molecular, el pI teórico, la composición de aminoácidos, la composición atómica, el coeficiente de extinción, etc.

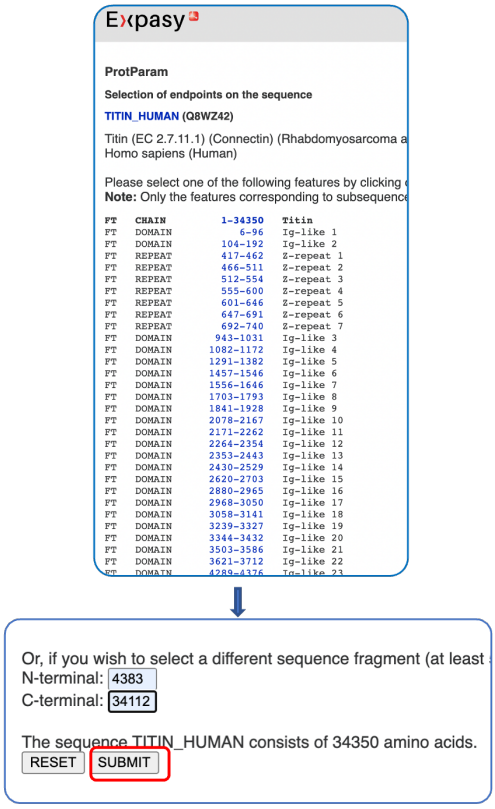


Los parámetros que se van a analizar son:

- Coeficiente de extinción: Muestra cuánta luz absorbe una proteína a una cierta longitud de onda y resulta muy útil en los estudios de espectrometría. *ProtParam* ofrece una estimación que, en todo caso, debe confirmarse experimentalmente.
- Coeficiente de inestabilidad: La proteína se encuentra estable cuando el valor es de 40, y cuando el valor mayor indica que la proteína puede ser inestable.
- Vida media: Es una predicción del tiempo que tarda la proteína completa en desaparecer después de ser sintetizada en la célula.

7. En el caso de que se ingrese el número identificador (Q8WZ42), se muestra una pantalla intermedia previa a la pantalla de resultados, en donde es posible seleccionar para el análisis de la secuencia completa o de los dominios funcionales.





8. Como resultado de la búsqueda específica para el análisis, se obtienen las propiedades fisicoquímicas del fragmento.

ProtParam

TITIN_HUMAN (Q8WZ42)

Titin (EC 2.7.11.1) (Connectin) (Rhabdomyosarcoma antigen MU-RMS-40.14)
Homo sapiens (Human)
The computation has been carried out on a user selected segment from position 4383 to position 34112 in this sequence of 34350 residues.

Considered sequence fragment:

```
      4381       4391       4401       4411       4421       4431
      |          |          |          |          |          |
 4381   PVIKRKIE PLEVALGHLA KFTCEIQSAP NVRFQWFKAG REIYESDKCS IRSSKYISSL  4440
 4441 EILRTQVVDC GEYTCKASNE YGSVSCTATL TVTEAYPPTF LSRPKSLTTF VGKAAKFICT  4500
 4501 VTGTPVIETI WQKDGAALSP SPNWRISDAE NKHILELSNL TIQDRGVYSC KASNKFGADI  4560
 4561 CQAELIIIDK PHFIKELEPV QSAINKKVHL ECQVDEDRKV TVTWSKDGQK LPPGKDYKIC  4620
 4621 FEDKIATLEI PLAKLKDSGT YVCTASNEAG SSSCSATVTV REPPSFVKKV DPSYLMLPGE  4680
 4681 SARLHCKLKG SPVIQVTWFK NNKELSESNT VRMYFVNSEA ILDITDVKVE DSGSYSCEAV  4740
```

Number of amino acids: 29730

Molecular weight: 3300936.55

Theoretical pI: 6.13

Amino acid composition: CSV format

Ala (A)	1762	5.9%
Arg (R)	1426	4.8%
Asn (N)	998	3.4%
Asp (D)	1521	5.1%
Cys (C)	460	1.5%
Gln (Q)	724	2.4%
Glu (E)	2714	9.1%
Gly (G)	1834	6.2%
His (H)	389	1.3%
Ile (I)	1787	6.0%
Leu (L)	1816	6.1%
Lys (K)	2579	8.7%
Met (M)	309	1.0%
Phe (F)	780	2.6%
Pro (P)	2279	7.7%
Ser (S)	2086	7.0%
Thr (T)	2187	7.4%
Trp (W)	430	1.4%
Tyr (Y)	873	2.9%
Val (V)	2776	9.3%
Pyl (O)	0	0.0%
Sec (U)	0	0.0%
(B)	0	0.0%
(Z)	0	0.0%
(X)	0	0.0%

Total number of negatively charged residues (Asp + Glu): 4235
Total number of positively charged residues (Arg + Lys): 4005

Atomic composition:

Carbon	C	147015
Hydrogen	H	234047
Nitrogen	N	39517
Oxygen	O	45069
Sulfur	S	769

Formula: $C_{147015}H_{234047}N_{39517}O_{45069}S_{769}$
Total number of atoms: 466417

Extinction coefficients:

Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm measured in water.

Ext. coefficient 3694520
Abs 0.1% (=1 g/l) 1.119, assuming all pairs of Cys residues form cystines

Ext. coefficient 3665770
Abs 0.1% (=1 g/l) 1.111, assuming all Cys residues are reduced

Estimated half-life:

The N-terminal of the sequence considered is P (Pro).

The estimated half-life is: >20 hours (mammalian reticulocytes, in vitro).
>20 hours (yeast, in vivo).
7 (Escherichia coli, in vivo).

Instability index:

The instability index (II) is computed to be 41.68
This classifies the protein as unstable.

Aliphatic index: 80.27

Grand average of hydropathicity (GRAVY): -0.470

Cuestionario y/o ejercicios complementarios

1. ¿Por qué es importante conocer el pI, el peso molecular, la solubilidad y otros datos de la proteína?

2. ¿Qué propiedades fisicoquímicas son importantes para la purificación de las proteínas?

Bibliografía

- Atwood, T., y Parry-Smith, D. (2002). *Introducción a la bioinformática*. Madrid, España: Pearson educación, S.A.
- Melo, V., y Cuamatzi, O. (2004). *Bioquímica de los procesos metabólicos*. Barcelona, España: Reverte, S.A.
- Peña, A., Arroyo, A., Gómez, A., y Tapia, R. (2018). *Bioquímica*. Ciudad de México, México: Limusa.
- Roldan, D. (2015). *Bioinformática, el ADN en un solo clic*. Bogotá, Colombia: Ediciones de la U. Stryer, L., Berg, J. y Tymokzko, J. (2013). Bioquímica. Madrid, España: Reverté, S.A.

Práctica 6

Modificaciones postraduccionales

MOTIF SCAN

Introducción

Las modificaciones postraduccionales son eventos de procesamiento que cambian las propiedades de una proteína mediante la ruptura proteolítica y/o la adición covalente de un grupo modificador, como acetilo, fosforilo, glicosilo y metilo, a uno o más aminoácidos. Estas juegan un papel clave en numerosos procesos biológicos al modificar significativamente la estructura y dinámica de las proteínas. Estas modificaciones desencadenan una amplia gama de comportamientos y características de las proteínas, incluida la función y el ensamblaje de las enzimas, la vida útil de las proteínas, las interacciones proteína-proteína, las interacciones célula- célula y célula-matriz, el tráfico molecular, la activación del receptor, la solubilidad de la proteína, plegamiento de proteínas y localización de proteínas.

Estas modificaciones no se producen en cualquier parte de la proteína, sino que vienen determinadas por una secuencia concreta de aminoácidos conocida como motivos proteicos, que son reconocidos por las enzimas encargadas de realizar estas modificaciones postraduccionales. Es posible encontrar estos patrones en las proteínas ya que están descritos y depositados en diferentes bases de datos.

MOTIF SCAN

Si una proteína tiene sitios susceptibles a presentar modificaciones postraduccionales se utiliza el programa *Motif Scan*, el cual compara la secuencia de aminoácidos con las bases de datos de motivos proteicos incluyendo "*prosite*", una de las bases más importantes. Como resultado se obtienen las coincidencias reportadas en las bases de datos y algunos otros detalles.

Objetivo general

- Identificar motivos en la secuencia proteica en donde se sugieren modificaciones postraduccionales mediante el programa *Motif Scan*.

Objetivos particulares

1. Reconocer las modificaciones postraduccionales de las proteínas mediante la base de datos *Motif Scan* con el fin de compararla con otras proteínas.

2. Identificar la secuencia de un motivo señalando la posición encontrada para compararla con otras proteínas.

Requerimientos para la práctica

- Dispositivo con conexión a Internet.

Nota: No es necesario descargar ningún programa para realizar esta práctica, se trabaja en línea.

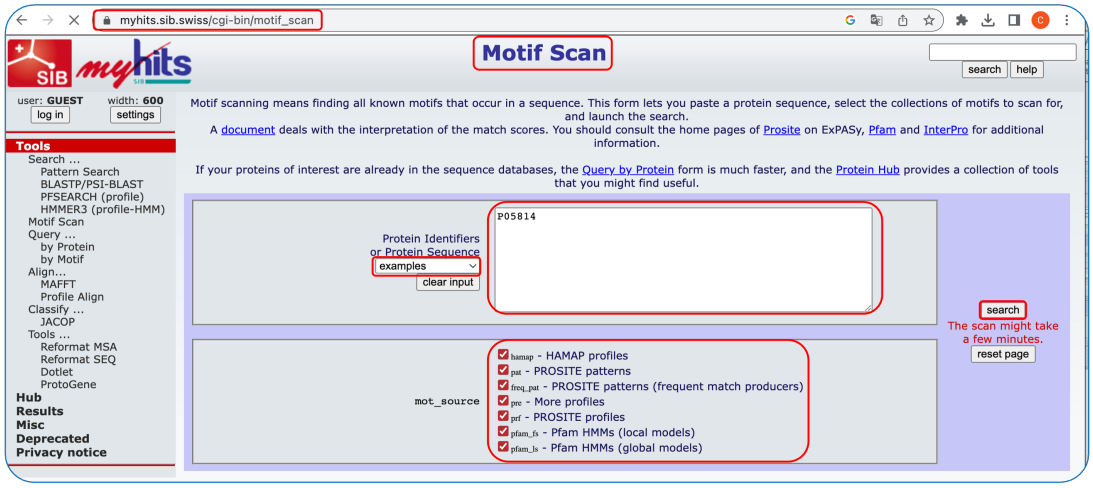
Procedimiento

1. Acceder a *Motif Scan* en la siguiente liga: https://myhits.isb-sib.ch/cgi-bin/motif_scan#.

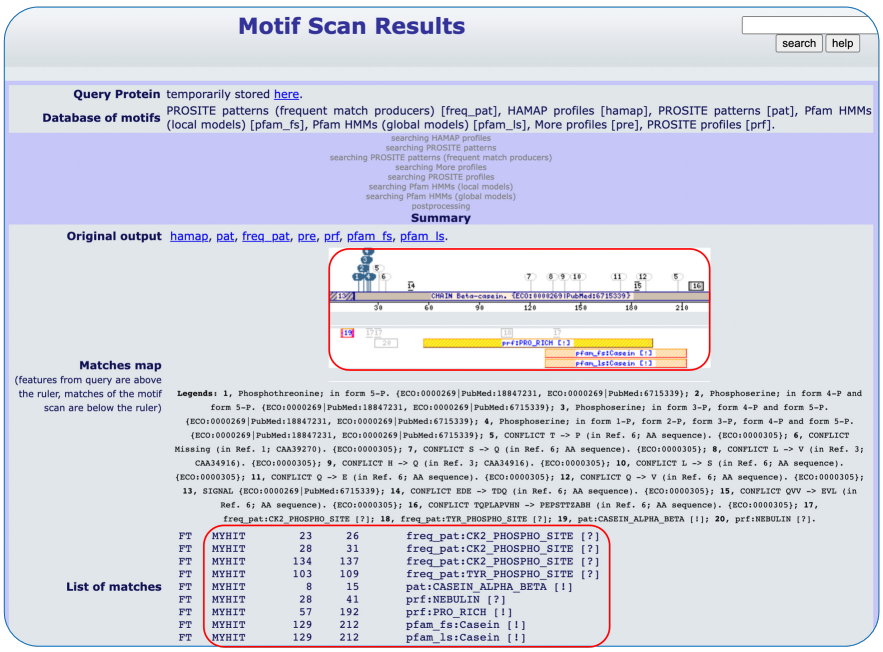
2. Se buscarán los motivos de *β-casein* de *Homo sapiens*, cuyo número identificador es P05814, para ello buscar la proteína en *Uniprot* (https://www.uniprot.org/) y copiar el *número identificador* (*Entry: Unique and stable entry identifier*). También se puede utilizar en este paso la secuencia en formato *FASTA* de la proteína.

3. Introducir el número identificador o la secuencia en formato *FASTA* de la proteína en el recuadro que se encuentra en la página principal.

4. Seleccionar todas las bases de datos en donde se buscarán las coincidencias en las secuencias proteicas reportadas con modificaciones postraduccionales. Dar clic en "*Search*".



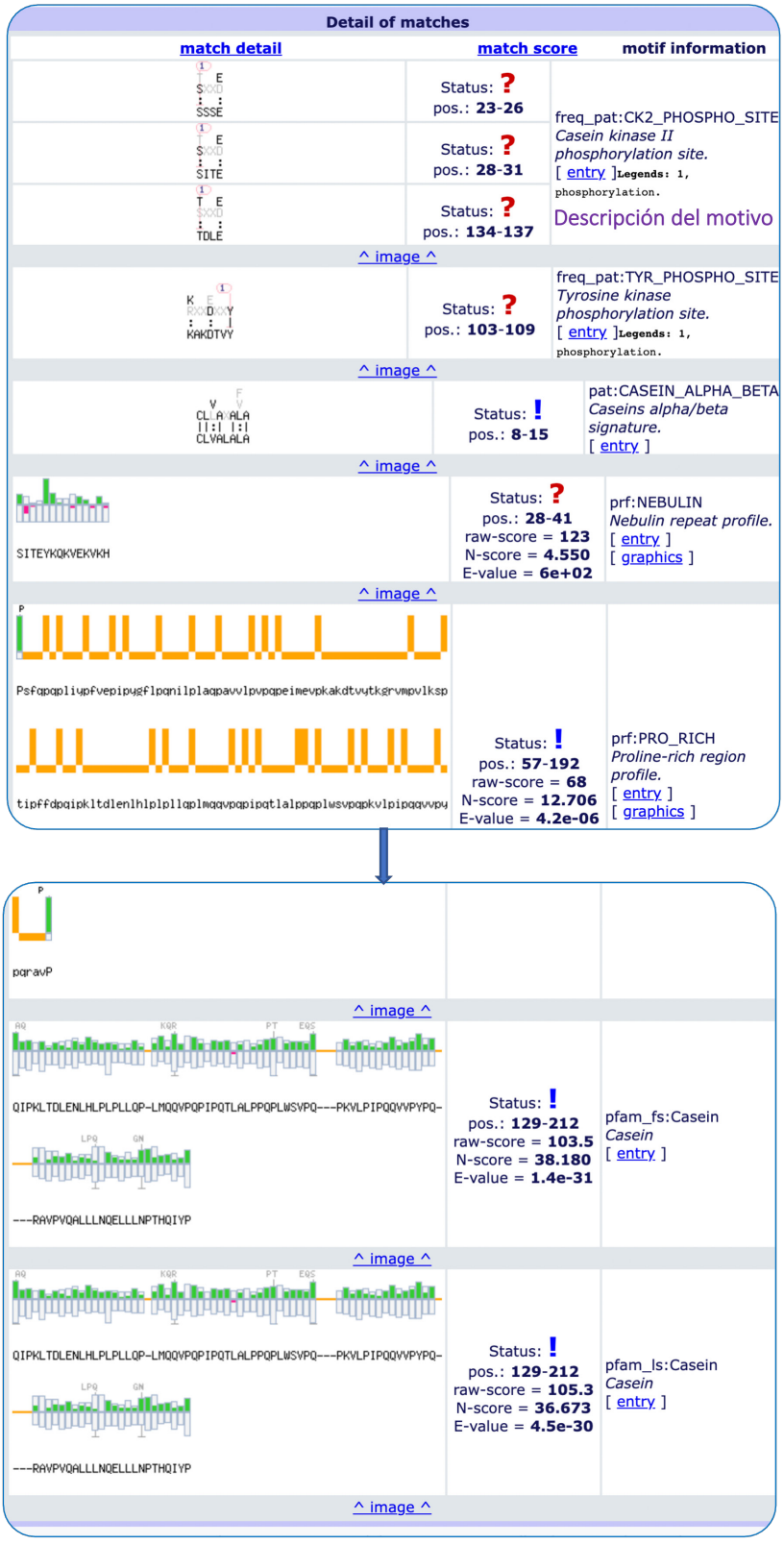
5. Como resultado se obtiene una representación gráfica en la cual se observa la secuencia proteica introducida en la parte superior y en la parte inferior las coincidencias obtenidas por la búsqueda en las diferentes bases de datos, seguida de una lista de las coincidencias.



6. En la imagen anterior se muestran los detalles de cada motivo y la posición, cuando se encuentra disponible nos permite obtener una imagen o esquema y la descripción del motivo, en este caso encontramos sitios ricos en prolina y sitios de fosforilación. La significancia se muestra mediante un código de coincidencias, en las que, según lo obtenido en las bases de datos, las secuencias coincidentes proporcionadas pueden clasificarse como verdaderos positivos y falsos positivos. Un verdadero positivo es una secuencia que comparte similitud con la secuencia de estudio porque ambos han evolucionado (divergido) de una secuencia ancestral común, aunque también puede atribuirse a veces a la convergencia evolutiva. Una secuencia se considera un falso positivo si la similitud observada es atribuible al azar. Debe enfatizarse que solo los argumentos biológicos pueden ser permitidos para decidir si una secuencia debe considerarse como un verdadero o falso positivo.

7. Según el código proporcionado por el programa:

    - **!**. Indica una coincidencia fuerte: Es poco probable que esta coincidencia sea un falso positivo.
    - **R**. Coincidencia rescatada: Esto se refiere principalmente a dominios que se sabe que se repiten y que es poco probable que aparezcan como una sola copia en una proteína.
    - **?**. Coincidencia cuestionable o débil: Se requieren evidencias biológicas adicionales para determinar el estado de esta coincidencia.
    - **!!**. Una fuerte coincidencia para un motivo específico de una familia de proteínas: Es muy poco probable que esta coincidencia sea un falso positivo, además, es muy probable que esta coincidencia pertenezca a la subfamilia.
    - **?!**. Corresponde a una coincidencia aceptada para un motivo específico de la familia: Es muy poco probable que esta coincidencia sea un falso positivo para el motivo, pero para determinar la asignación familiar se requiere de evidencias biológicas adicionales.
    - **??**. Coincidencia cuestionable o débil para un motivo específico.
    - **NA**: No disponible.
    - **Valor E**: Es el número de coincidencias con una puntuación igual o mayor que la puntuación observada que se espera que ocurra por casualidad. En otras palabras, el valor E proporciona una estimación del número de falsos positivos.


Detail of matches
match detail
match score
motif information
Status: ?
pos.: 23-26
freq_pat:CK2_PHOSPHO_SITE
Casein kinase II phosphorylation site.
[entry] Legends: 1, phosphorylation.
Status: ?
pos.: 28-31
Status: ?
pos.: 134-137
Descripción del motivo
^ image ^
Status: ?
pos.: 103-109
freq_pat:TYR_PHOSPHO_SITE
Tyrosine kinase phosphorylation site.
[entry] Legends: 1, phosphorylation.
^ image ^
Status: !
pos.: 8-15
pat:CASEIN_ALPHA_BETA
Caseins alpha/beta signature.
[entry]
^ image ^
SITEYKQKVEKVKH
Status: ?
pos.: 28-41
raw-score = 123
N-score = 4.550
E-value = 6e+02
prf:NEBULIN
Nebulin repeat profile.
[entry]
[graphics]
^ image ^
Psfqpqpliypfvepipygflpqnilplaqpavvlpvpqpeimevpkakdtvytkgrvmpvlksp
tipffdpqipkltdlenlhlplpllqplmqqvpqpipqtlalppqplwsvpqpkvlpipqqvvpy
Status: !
pos.: 57-192
raw-score = 68
N-score = 12.706
E-value = 4.2e-06
prf:PRO_RICH
Proline-rich region profile.
[entry]
[graphics]
pqravP
^ image ^
QIPKLTDLENLHLPLPLLQP-LMQQVPQPIPQTLALPPQPLWSVPQ---PKVLPIPQQVVPYPQ-
---RAVPVQALLLNQELLLNPTHQIYP
Status: !
pos.: 129-212
raw-score = 103.5
N-score = 38.180
E-value = 1.4e-31
pfam_fs:Casein
Casein
[entry]
^ image ^
QIPKLTDLENLHLPLPLLQP-LMQQVPQPIPQTLALPPQPLWSVPQ---PKVLPIPQQVVPYPQ-
---RAVPVQALLLNQELLLNPTHQIYP
Status: !
pos.: 129-212
raw-score = 105.3
N-score = 36.673
E-value = 4.5e-30
pfam_ls:Casein
Casein
[entry]
^ image ^

8. De acuerdo con los resultados obtenidos en el paso anterior, completa el siguiente cuadro:

Descripción del motivo proteico	Secuencia del motivo	Posición en la secuencia de la proteína	"*Score*" i=fuerte, !=débil

Cuestionario y/o ejercicios complementarios

1. Describir la función de los motivos encontrados y las modificaciones postraduccionales encontradas en la secuencia.

2. Investigar otros programas para encontrar posibles modificaciones postraduccionales.

3. Repetir el ejercicio con NFKB1 (*Homo sapiens*), con identificador: P19838. ¿Qué resultados se obtienen? Realizar un cuadro con los resultados y describir la función de los motivos encontrados y las posibles modificaciones postraduccionales.

Bibliografía

- Atwood, T., y Parry-Smith, D. (2002). *Introducción a la bioinformática*. Madrid, España: Pearson educación, S.A.
- Oliva, R., y Vidal, M. (2006). *Genoma humano. Nuevos avances en investigación, diagnóstico y tratamiento*. Barcelona, España: UBe Barcelona.
- Roldan, D. (2015). *Bioinformática, el ADN en un solo clic*. Bogotá, Colombia: RA-MA ediciones de la U.
- Charpilloz C., Veuthey, A-L., et al. (2014). *Motifs tree: a new method for predicting post-translational modifications*. Bioinformatics. 30:1974-1982. https://doi.org/10.1093/bioinformatics/btu165

Práctica 7

Búsqueda de dominios funcionales

INTERPRO

Introducción

Algunas regiones de la estructura de una proteína se les conocen como dominios. Existen tres clases principales de dominios proteicos: funcional, estructural y topológico.

El dominio funcional es una región de una proteína que muestra una actividad particular y es conservada aun cuando se aísle del resto de la molécula. A este respecto una región específica de una proteína puede ser responsable de la actividad catalítica, por ejemplo, el dominio cinasa que adiciona en forma covalente un grupo fosfato a otra proteína o un dominio de unión a DNA o un dominio de unión a la membrana. Un dominio funcional con frecuencia se identifica en forma experimental al reducir una proteína a un fragmento activo más pequeño con ayuda de las proteínas proteasas, enzimas que escinden uno o más enlaces peptídicos en un polipéptido diana.

El dominio estructural es una región de aproximadamente 40 aminoácidos, dispuesta como una estructura única es estable y distintiva, con frecuencia abarca una o más estructuras secundarias.

El dominio topológico se refiere a las regiones de las proteínas definidas por las relaciones espaciales distintivas con respecto al resto de la proteína, por ejemplo, algunas proteínas asociadas con la membrana de la superficie celular pueden tener una parte que se extiende hacia el espacio citoplasmático y una parte embebida en la membrana fosfolipídica y otra parte que se extiende hacia el exterior, espacio extracelular.

Plataforma interpro (clasificación de familias de proteínas)

InterPro provee un análisis funcional de las proteínas mediante la clasificación en familias y predicción de dominios y sitios importantes. Para clasificar las proteínas, usa modelos predictivos conocidos como firmas ("*Signatures*") proporcionadas por diferentes bases de datos que conforman *InterPro*. Se combina la información de las diferentes bases de datos en una sola, funcionando como una herramienta de diagnóstico.

Objetivo general

- Identificar dominios funcionales en las secuencias proteicas mediante una base de datos con el fin de encontrar fragmentos conservados para inferir la posible función.

Objetivos particulares

1. Identificar los dominios funcionales reportados mediante una plataforma con acceso a base de datos sobre proteínas con el fin de encontrar coincidencias con otras familias de proteínas.

2. Reconocer con base en la secuencia, los sitios de unión de la proteína para inferir la función molecular y el proceso biológico en el que está involucrada.

Requerimientos para la elaboración de la práctica

- Dispositivo con conexión a Internet.

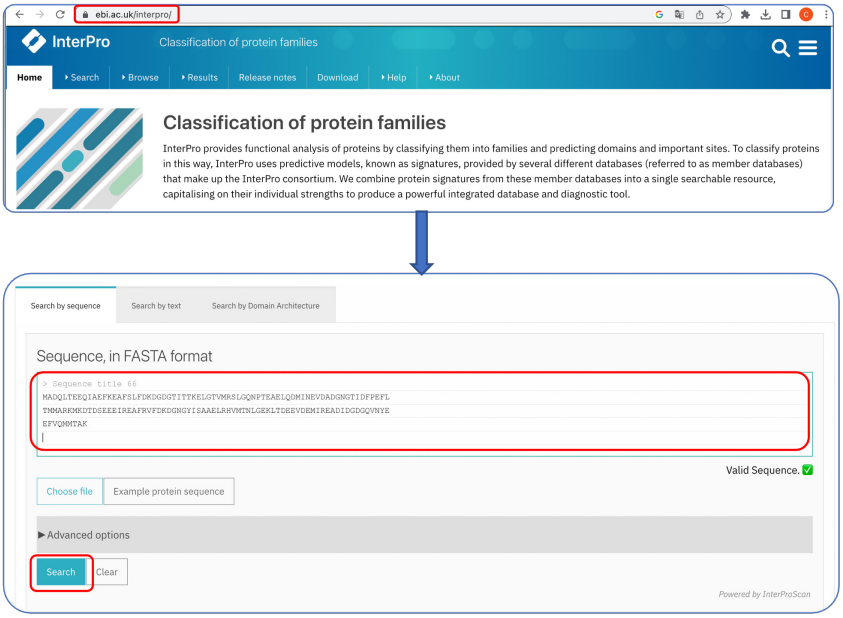
Nota: No es necesario descargar ningún programa para realizar esta práctica, se trabaja en línea.

Procedimiento

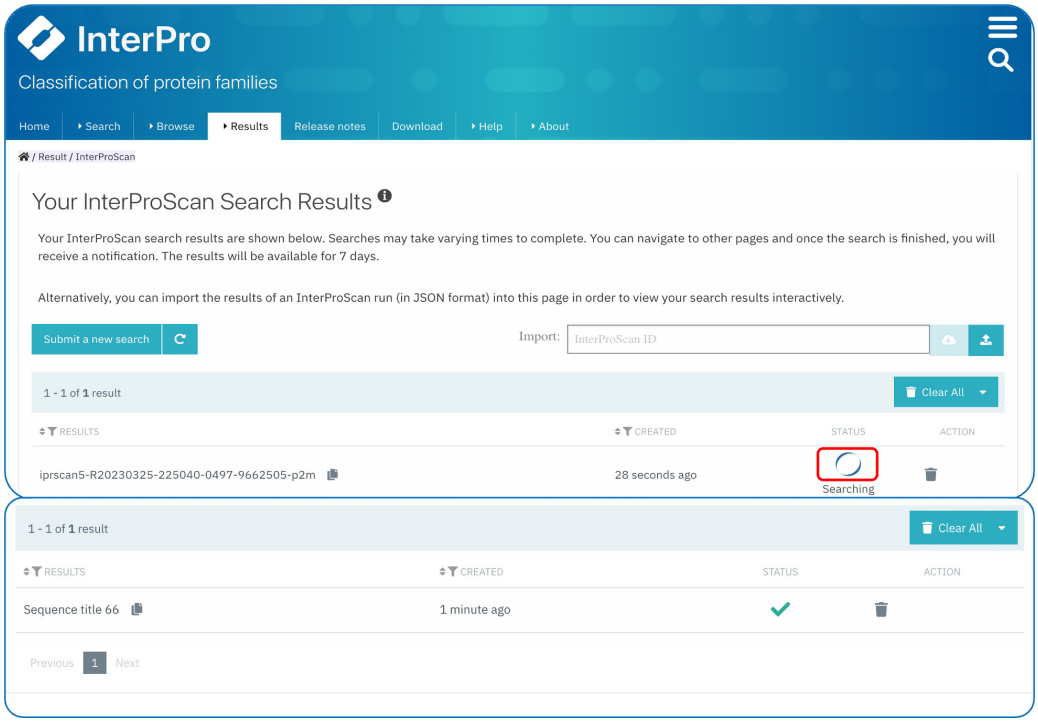
1. Esta práctica se llevará a cabo utilizando la secuencia de calmodulina humana >AAD45181.1 *calmodulin* [*Homo sapiens*]. Obtener el formato *FASTA* en *NCBI/Protein*.

2. Abrir la plataforma *InterPro* en el siguiente enlace https://www.ebi.ac.uk/interpro/

3. Una vez abierta la página principal de *InterPro*, colocar la secuencia de la proteína en formato *FASTA* en la ventana. Dar clic en "*Search*".



4. El procesamiento de datos puede ser lento (se muestra el estatus, espere un momento).



5. Una vez que el procedimiento ha finalizado, aparecerá la página de resultados, que dispone de una sección de filtrado en donde es posible especificar el tipo de resultados que se desee según una serie de criterios. El más importante de ellos es el tipo de entrada ("*Entry*"). A cada entrada de *InterPro* se le asigna uno de los siguientes tipos, que permiten inferir cuando una proteína coincide con una entrada:

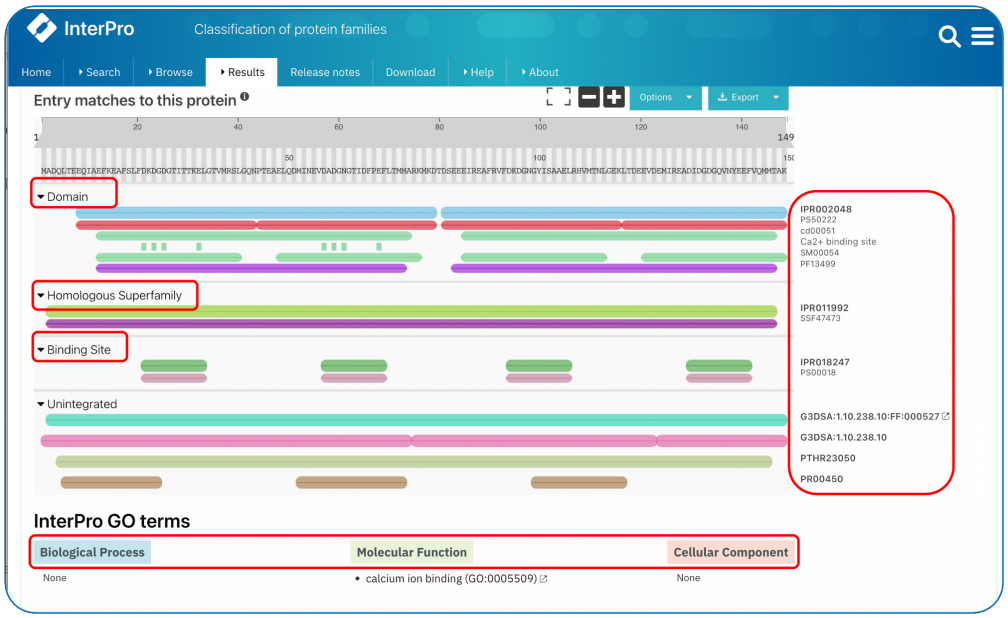
    - Familia ("*Family*"): Una familia de proteínas que comparten una evolución común, lo que indica que desempeñan funciones relacionadas o que tienen una estructura secundaria o terciaria similar.

    - Dominio ("*Domain*"): Los dominios pueden existir en una amplia variedad de contextos biológicos y se caracterizan por una estructura, función o fragmento de secuencia.

    - Súper familia homóloga ("*Homologous Superfamily*"): Es un grupo de proteínas que comparten un origen evolutivo común indicado por la similitud estructural.

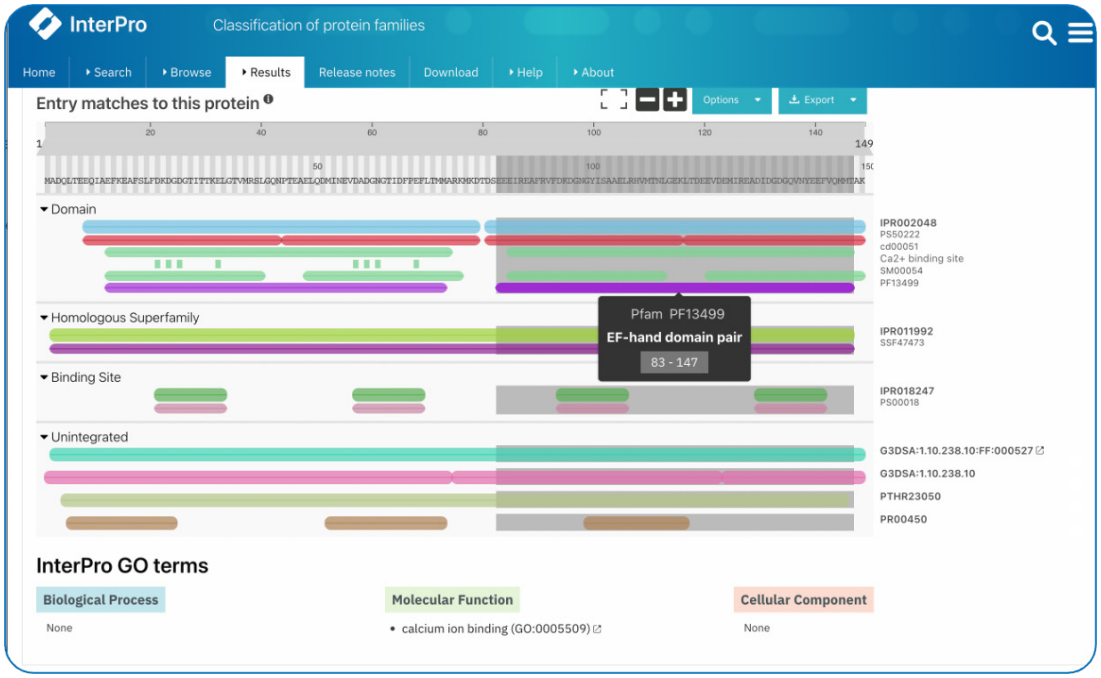
    - Sitios de unión ("*Binding site*"): Un sitio de unión es una secuencia en una proteína en donde se une un ligando.

    - *Unintegrated*: Ligandos cuya unión es transitoria, como, por ejemplo, el ion calcio.

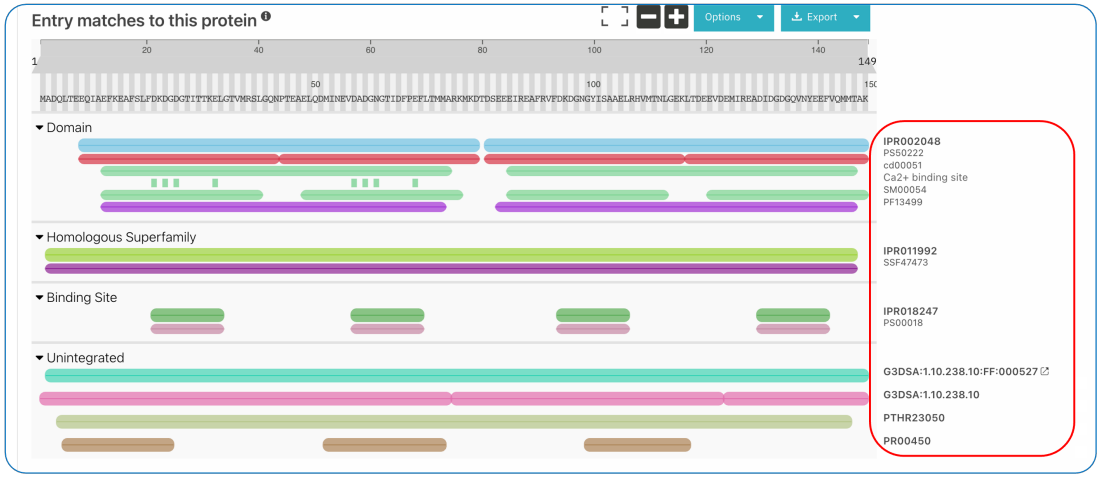


El resultado muestra que la secuencia pertenece a la familia calmodulin (IPR039030), en la que se han encontrado coincidencias en los dominios IPR002048, PS50222, SM00054, Cd00051, sitio de unión a Ca2+ ("Binding Site"), PF13499, mostradas en el recuadro marcado en rojo del lado derecho. Abajo del cuadro se muestran algunas características de los dominios.

6. Al colocar el cursor sobre cada una de las barras de las diferentes entradas se observan las coincidencias reportadas en las bases de datos, como se muestra en el ejemplo del recuadro rojo.



7. Es posible exportar los resultados en varios formatos y consultar las bases de datos biológicas de las que se ha obtenido la información. Para ello dar clic en las coincidencias de cada entrada, que se reportan en el lado derecho de la pantalla (recuadro en color rojo).



Cuestionario y/o ejercicios complementarios

1. ¿En qué parte de la estructura de la proteína se pueden encontrar dominios funcionales? Justificar la respuesta.
2. ¿Qué función tienen los dominios funcionales encontrados en la superfamilia?
3. ¿Qué es un sitio de unión proteico y que importancia biológica tiene?

Bibliografía

- Atwood, T., y Parry-Smith, D. (2002). *Introducción a la bioinformática*. Madrid, España: Pearson educación, S.A.
- Capel, J., y Yuste, F. (2016). *Manual de prácticas de bioinformática*. Andalucía España: Almería.
- Lodish, H., Berk, A., Kaiser, C., Kriegeer, M., Bretscher, A., Ploegh, H., Amon, A., y Scott, M. (2016). *Biología celular y molecular*. Buenos Aires, Argentina: Médica Panamericana.
- Oliva, R., y Vidal, J. (2006). *El genoma humano: nuevos avances en investigación, diagnóstico y tratamiento*. Barcelona, España: UBe.
- Roldan, D. (2015). *Bioinformática, el ADN en un solo clic*. Bogotá, Colombia: Ediciones de la U. Stryer, L., Berg, J., y Tymokzko, J. (2013). Bioquímica. Madrid, España: Reverté, S.A.

Práctica 8

Diseño, predicción y comparación de estructura secundaria de proteínas

PSIPRED

Introducción

El segundo nivel en la estructura jerárquica de las proteínas es la estructura secundaria, está constituida por un esqueleto polipeptídico que no asume una estructura tridimensional al azar, sino que, en lugar de ello, forma acomodos regulares de aminoácidos que se localizan cercanos entre sí, estos acomodos se denominan estructura secundaria del polipéptido. Un polipéptido puede contener múltiples tipos de estructura secundaria en diversas porciones de la cadena, dependiendo de la secuencia de los aa. Las estructuras secundarias se estabilizan a través de puentes de hidrógeno entre los átomos del esqueleto peptídico.

Las principales estructuras secundarias son:

- Conformación α hélice: Es una estructura rígida, en espiral que gira hacía la derecha, consta de un esqueleto polipeptídico central estrechamente compactado y enrollado mediante las cadenas laterales de los L-aminoácidos extendiéndose hacia afuera desde el eje central para evitar la interferencia estérica entre sí.

- Conformación lámina β o β plegada: El esqueleto peptídico de la lámina β está casi totalmente extendido, se pueden formar puentes de hidrógeno entre diferentes partes de una sola cadena que se pliega sobre sí misma (enlaces intracadena) o entre diferentes cadenas (enlaces intercadena). Los puentes de hidrógeno entre cadenas dan pie a una estructura repetida en zigzag, de ahí también el nombre de "lámina plegada". Por su aspecto de "pliegues" se les llama β plegadas.

- Conformación doblez β: Esta conformación invierte la dirección de una cadena polipeptídica y la ayuda a tomar una forma compacta y globular. Por lo general, se encuentran en la superficie de las moléculas proteicas y con frecuencia incluyen residuos cargados. Reciben este nombre porque conectan cadenas sucesivas de láminas β antiparalelas.

PSIPRED

PSIPRED es un servidor de predicción de la estructura de proteínas, el cual permite a los usuarios enviar una secuencia de proteínas para realizar una predicción y recibir los resultados tanto textualmente como por correo electrónico de manera gráfica de Internet. El usuario puede seleccionar uno de los tres métodos de predicción para aplicar a la secuencia: *PSIPRED*, un método de predicción de estructura secundaria de alta precisión; *MEMSAT 2*, una versión de un método de predicción de topología transmembrana ampliamente utilizado; o bien *GenTHREADER*, un método de reconocimiento de pliegues basado en el perfil de secuencia.

Objetivo general

- Predecir el tipo de estructura secundaria en proteínas globulares mediante el servidor *PSIPRED* con el fin de asociarlo con la función.

Objetivos particulares

1. Identificar la(s) estructura(s) secundaria(s) de las proteínas mediante una base de datos para inferir el tipo de conformación e interacciones de la macromolécula.

2. Reconocer los tipos de aminoácidos que integran la estructura secundaria de las proteínas mediante el análisis de *PSIPRED* para inferir la conformación espacial.

3. Comparar proteínas en lo que se refiere a la conformación mediante el resultado obtenido del servidor *PSIPRED* con el objetivo de reconocer la función.

Requerimientos para la práctica

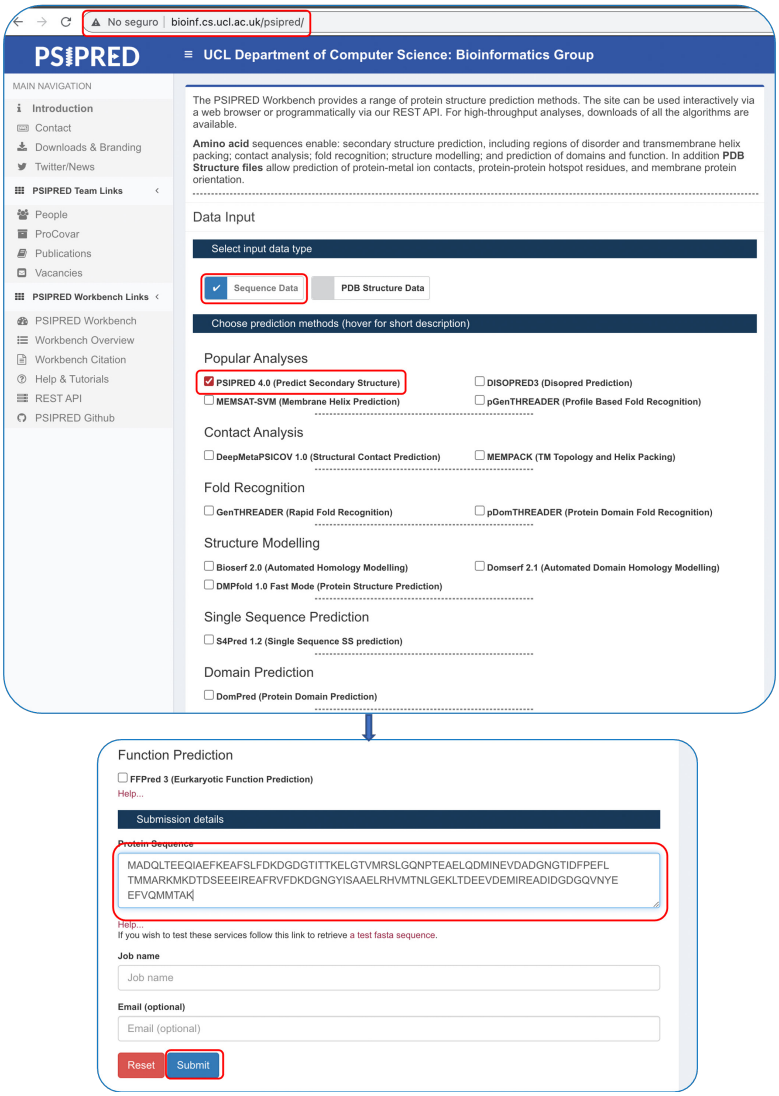
- Dispositivo con conexión a Internet.

Nota: No es necesario descargar ningún programa para realizar esta práctica, se trabaja en línea.

Procedimiento

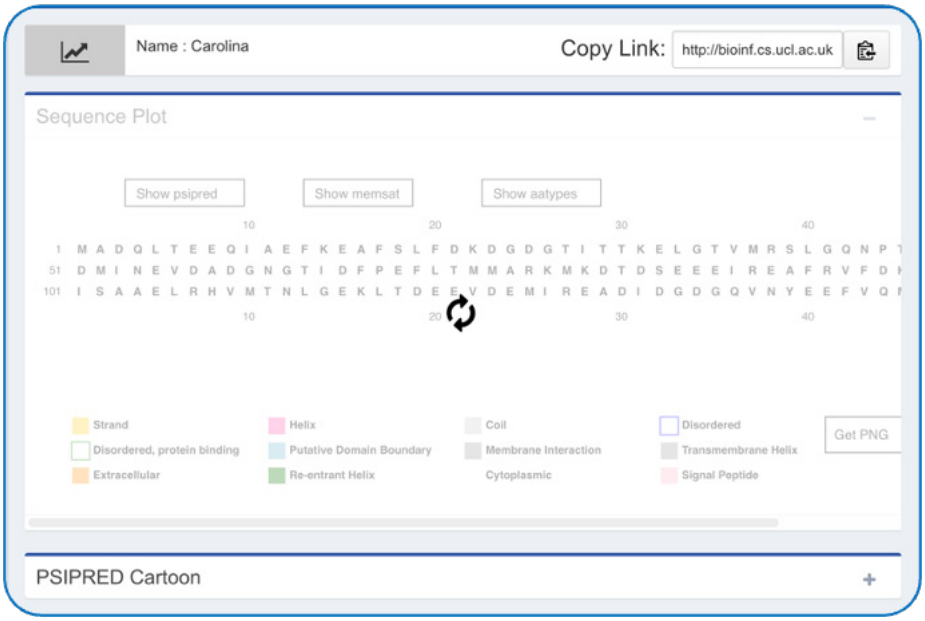
1. Obtener el formato *FASTA* de las proteínas *Calmodulin* (*Homo sapiens*), y *Green- fluorescent protein* (*Aequorea victoria*) en *NCBI/protein/FASTA*. Se realizará el análisis por separado de cada una.

2. Acceder a la herramienta *PSIPRED* en el siguiente enlace: http://bioinf.cs.ucl.ac.uk/psipred/, pegar la secuencia en formato *FASTA* de la primera proteína en la ventana indicada en rojo "*Submission details*".

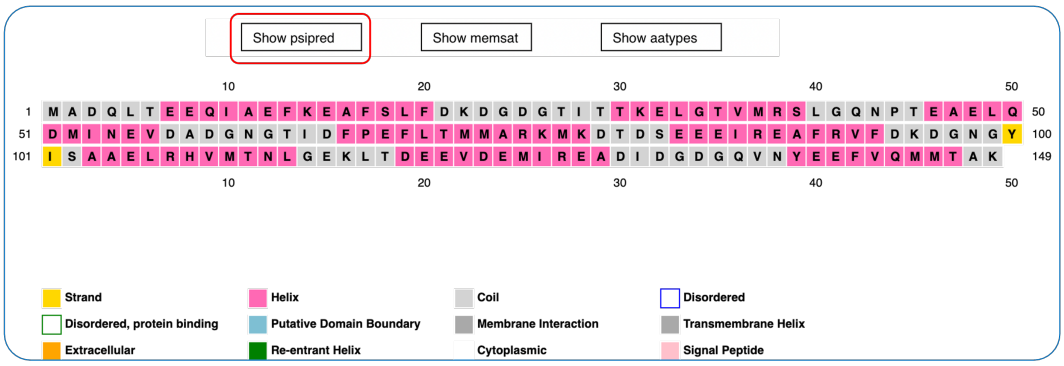


3. Una vez introducida la secuencia, dar clic en "*Submit*". *PSIPRED* brinda la opción de recibir los resultados vía correo electrónico, para ello introducir la dirección en la sección correspondiente.

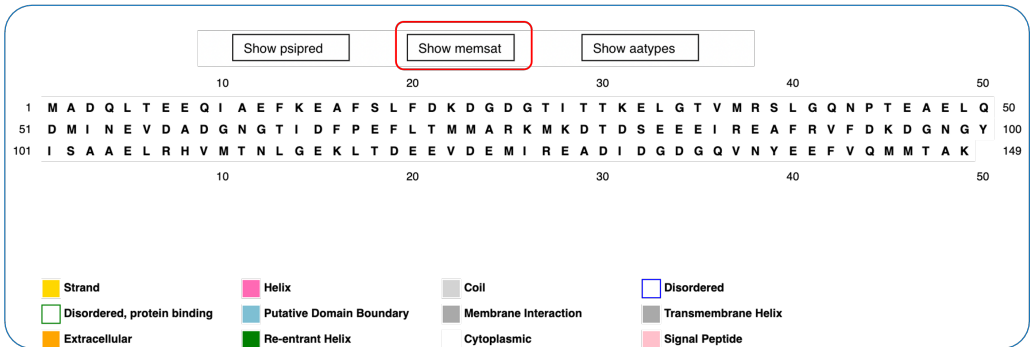
4. Aparece una pantalla mientras se está realizando el proceso de la información, puede tardar unos minutos.



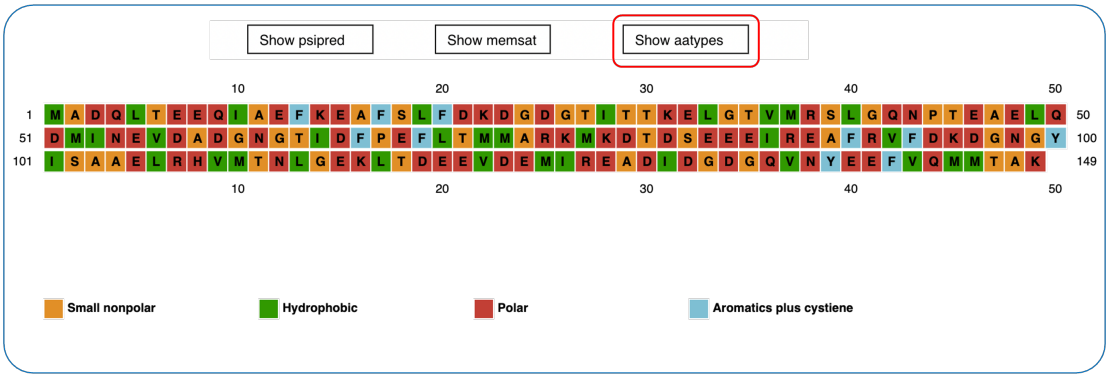
5. Una vez finalizado el análisis, se muestra la página de resultados, en la que se distinguen tres pestañas. En la primera se visualiza el mapa de la estructura secundaria que consiste en la secuencia de aminoácidos introducida, se muestran los aminoácidos marcados con un código de colores que indica cual es la posición que adoptan en el espacio.



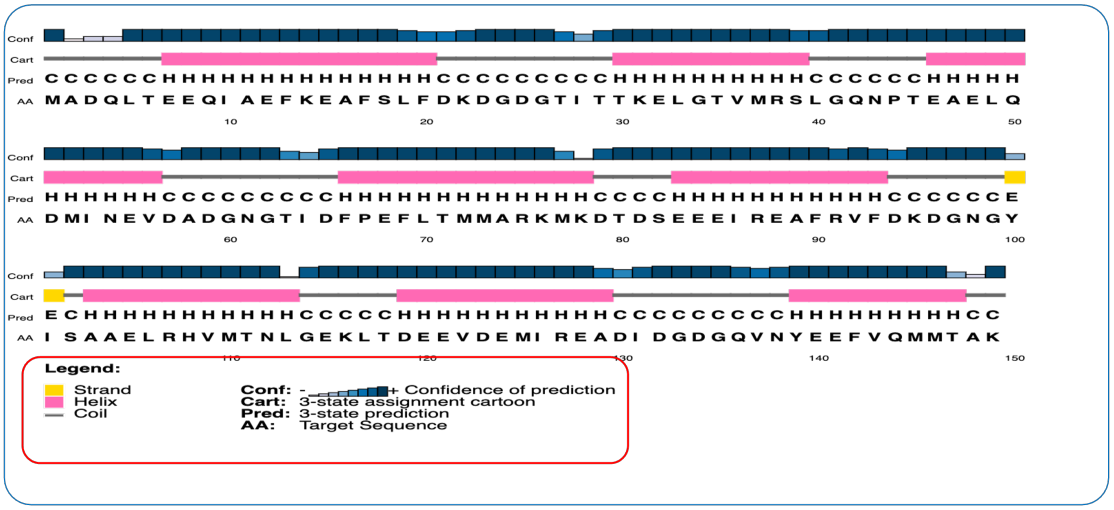
6. Dar clic en la segunda pestaña "*Show memsat*", *PSIPRED* muestra el mapa de la estructura secundaria representando en cuadros de colores, así como la posición que adopta cada sección de la proteína en el espacio celular.



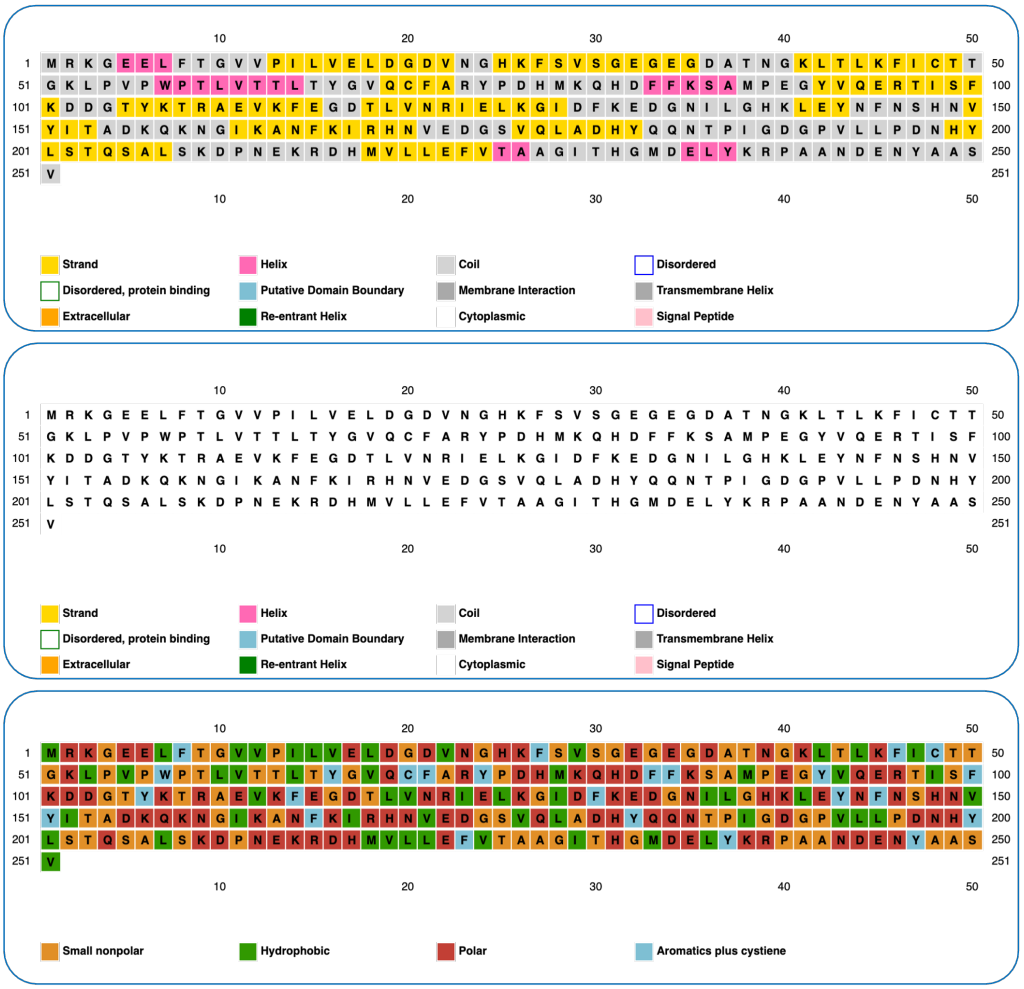
7. Dar clic en la tercera pestaña "*Show aatypes*" en esta sección se visualiza el mapa de cada aminoácido, se muestra con diferentes colores dependiendo de la carga.

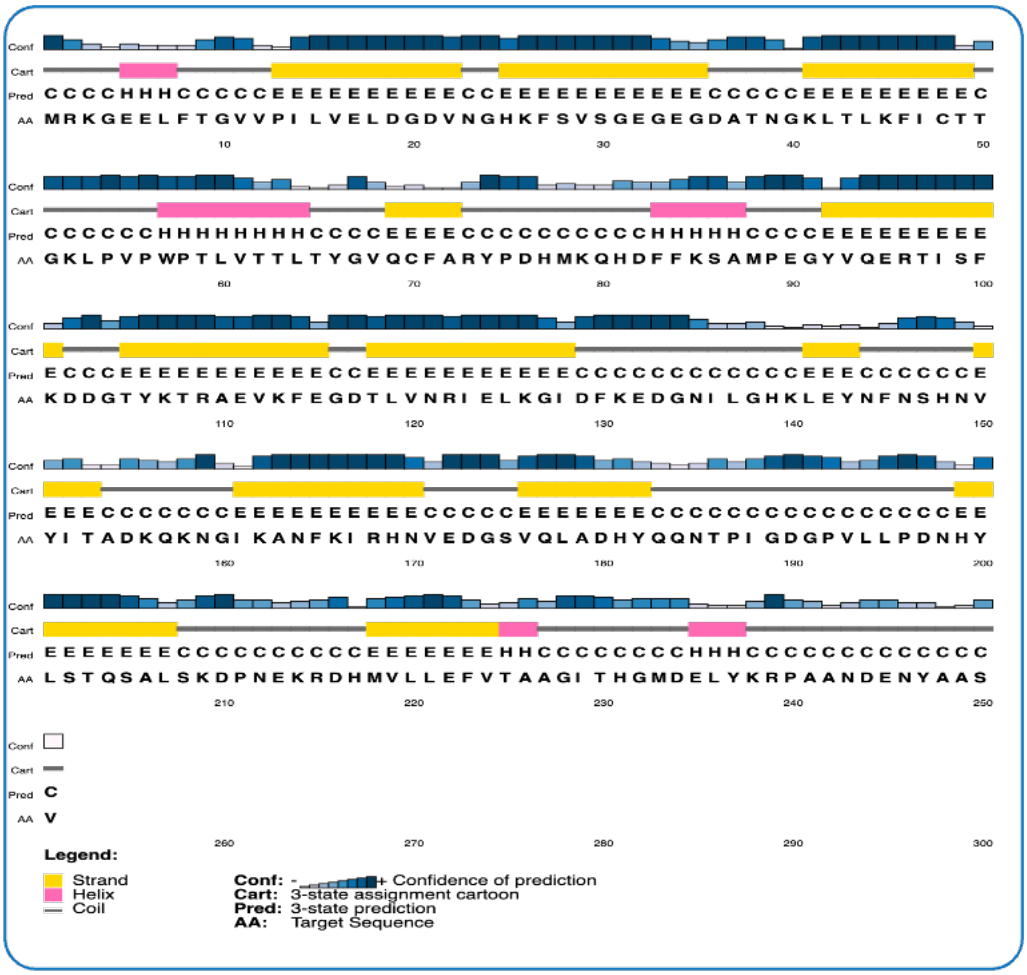


8. Dar clic en "*PSIPRED Cartoon* (+)" en esta sección se muestra la representación gráfica de la proteína en cuatro hileras, la primera corresponde a la de fiabilidad ("*Conf*") que indica el nivel de fiabilidad de la predicción para cada posición. La segunda muestra el tipo de configuración y consiste en una sucesión de hélices (H), plegamiento β (E) o giro β (C). La tercera línea ("*Pred*") asigna con letras el tipo de configuración. Por último, se visualiza la posición de los aminoácidos de la proteína.



9. Los resultados de la predicción de la estructura de la proteína *Green fluorescent* se muestran a continuación. Comparar y discutir los resultados obtenidos de la predicción de ambas proteínas.





10. Es posible descargar las imágenes dando clic en "*Get PNG*" o "*Get SVG*" dependiendo el formato deseado, así como el informe completo en distintos formatos a través de los enlaces que aparecen en la tercera pestaña.

Cuestionario y/o ejercicioscomplementarios

1. Realizar una correlación entre los tipos de configuración de la estructura secundaria y las cargas de los aminoácidos.
2. ¿Por qué el triptófano, la tirosina y la fenilalanina se encuentran a menudo en las láminas β?
3. ¿Qué tipo de aminoácidos se encuentran preferentemente en la α hélice?

Bibliografía

- Atwood, T., y Parry-Smith, D. (2002). *Introducción a la bioinformática*. Madrid, España: Pearson educación, S.A.
- Lodish, H., Berk, A., Kaiser, C., Kriegeer, M., Bretscher, A., Ploegh, H., Amon, A., y Scott, M. (2016). *Biología celular y molecular*. Buenos Aires, Argentina: Médica Panamericana.
- Macarulla, M., y Goñi, M. (2013). *Biomoléculas*. Barcelona, España: Reverté, S.A.
- Roldan, D. (2015). *Bioinformática, el ADN en un solo clic*. Bogotá, Colombia: Ediciones de la U.

Práctica 9

Modelado por homología: estructura terciaria

SWISS-MODEL

Introducción

La estructura terciaria es el siguiente nivel de complejidad en el plegamiento de las proteínas y se refiere a su forma tridimensional. Las proteínas se pliegan de tal manera que logran su máxima estabilidad y/o el estado de menor energía. Aunque la forma tridimensional de una proteína puede parecer irregular y aleatoria, está formada por diversas fuerzas que la estabilizan debido a las interacciones de enlaces entre los grupos de la cadena lateral de los aminoácidos, dentro de las que se encuentran: puentes disulfuro, interacciones iónicas, enlaces o uniones por puente de hidrógeno y fuerzas de van der Waals.

SWISS-MODEL

SWISS-MODEL es un servidor automatizado con interfaces Web fáciles de generar modelos confiables sin la necesidad de paquetes de software complejos, se especializa en estructuras de proteínas en tercera dimensión (3D), es muy sencillo de usar, totalmente automático y su fiabilidad es comparable a programas complejos. La opción de "modelo aproximado" comprueba si hay una estructura lo suficientemente homóloga para poder hacer el modelado, si existe, genera un primer modelo y datos que son necesarios para realizar un modelo refinado.

Objetivo general

- Predecir la estructura terciaria mediante el servidor SWISS-MODEL a través de una secuencia conocida con el fin de obtener la estructura tridimensional terciaria.

Objetivos particulares

1. Obtener la conformación tridimensional mediante un servidor a partir de una secuencia de aminoácidos.
2. Identificar por similitud las secuencias de las proteínas de acuerdo con su localización subcelular para comparar la estructura-función.
3. Buscar y alinear estructuras mediante herramientas bioinformáticas con el fin de conocer su homología.

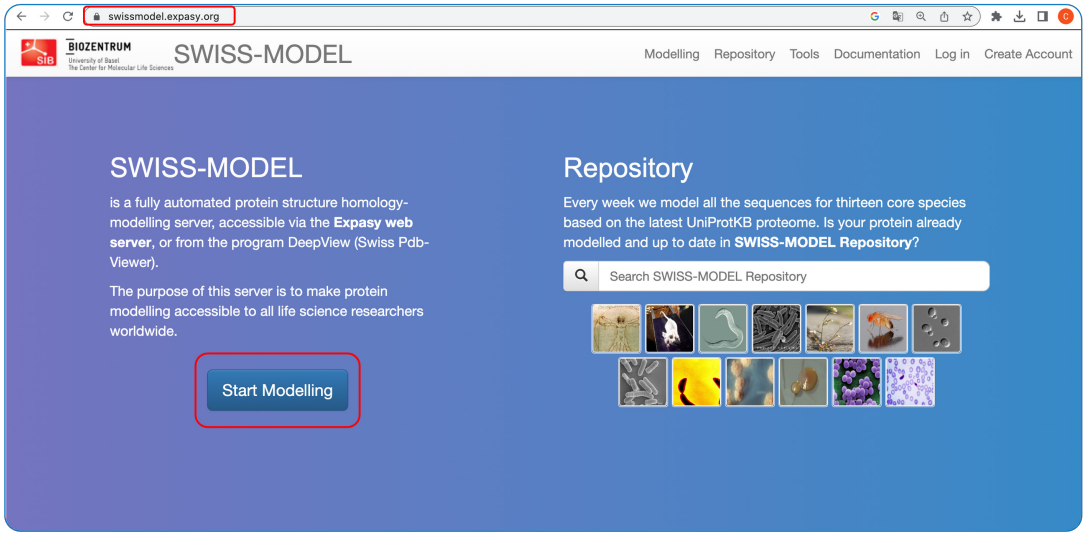
Requerimientos para la práctica

- Dispositivo con conexión a Internet.

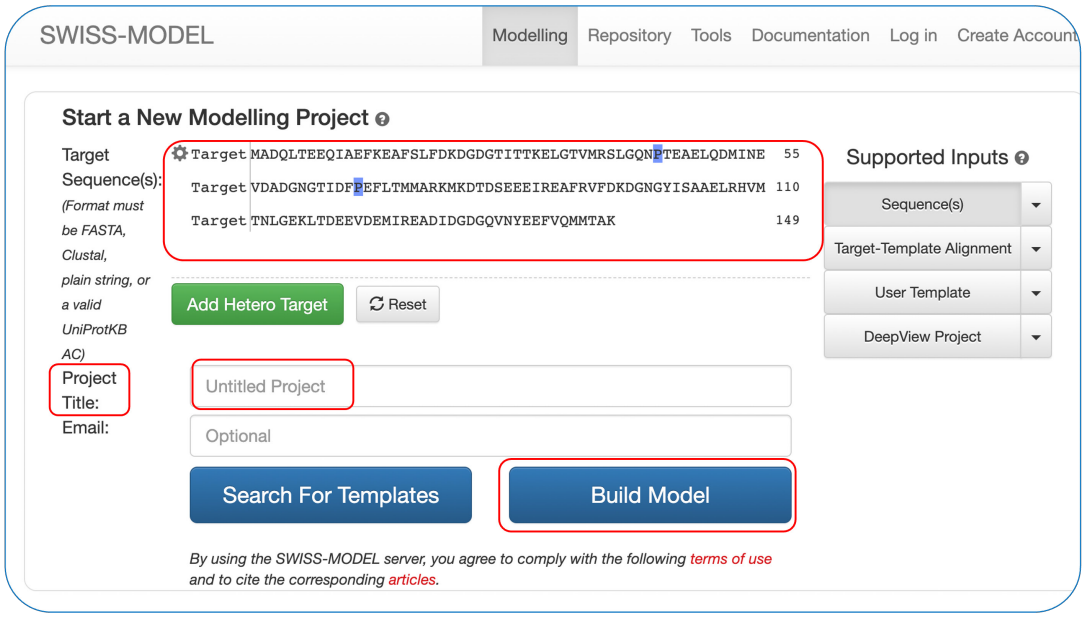
Nota: No es necesario descargar ningún programa para realizar esta práctica, se trabaja en línea.

Procedimiento

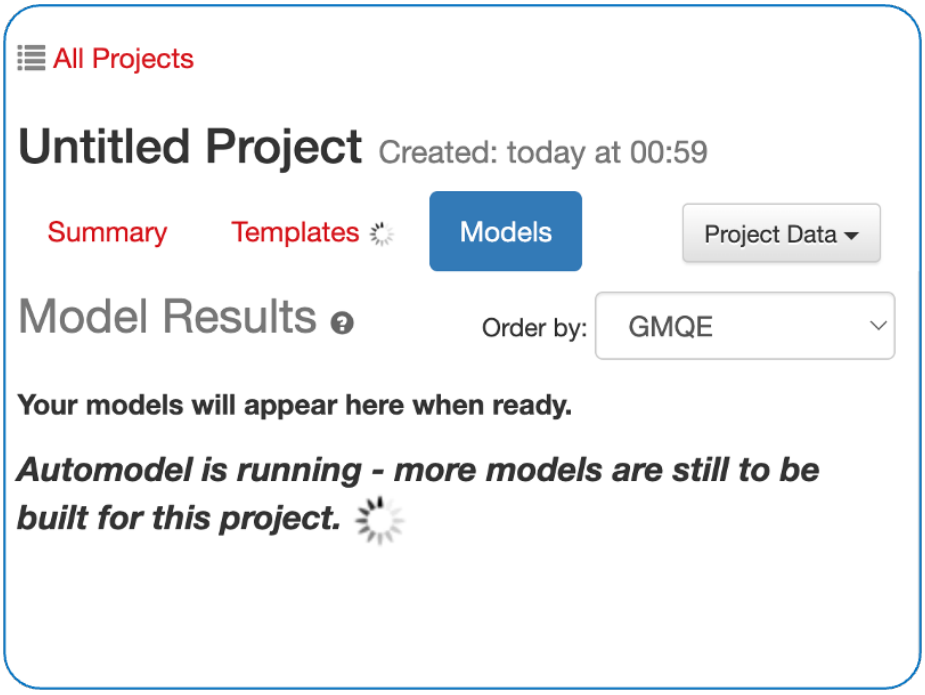
1. Esta práctica se realizará con las proteínas *Calmodulin* (*Homo sapiens*) y *Green fluorescent protein* (*Aequorea victoria*) por separado. Acceder a *NCBI/protein/FASTA*. Acceder a *SWISS MODEL* en la siguiente liga http://swissmodel.expasy.org/. En la página principal, dar clic en el botón "*Start Modelling*".



2. Introducir la secuencia de la proteína *calmodulin*. Agregar el nombre del proyecto y hacer clic en "*Build Model*".



3. El proceso puede tardar unos minutos.

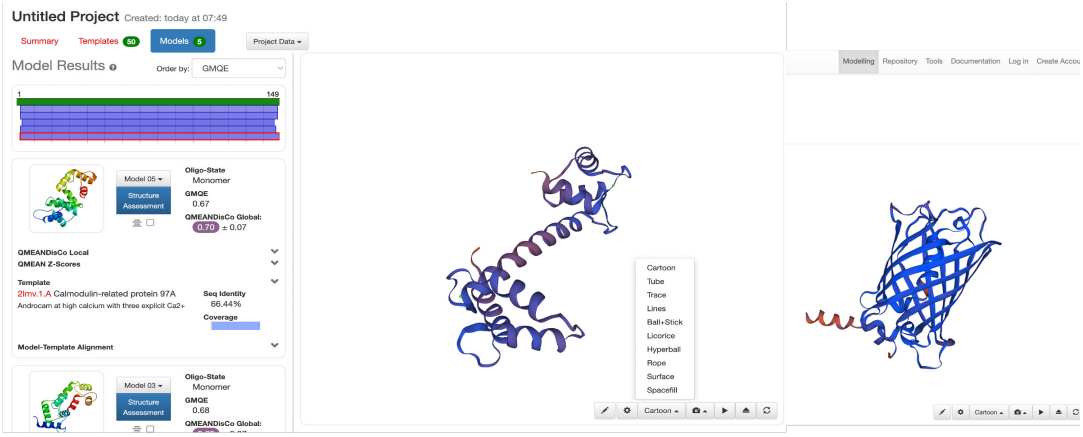


4. Los resultados muestran la predicción de la estructura terciaria tridimensional de la proteína, además de los parámetros de búsqueda.

5. En la parte inferior se muestra el alineamiento de la secuencia introducida para obtener el modelo. Del lado derecho se muestra la representación gráfica de la proteína en 3D, hacia abajo se encuentran las opciones "*NGL*" y "*Cartoons*" en donde podemos elegir el tipo de trazo (tubos, líneas, entre otros). Además, se muestra el icono de cámara, el cual permite tomar una fotografía de la proteína en 3D y el icono "*Play*" en donde podemos hacer rotar la imagen 3D, también puede girarse en el espacio utilizando el ratón/cursor.

6. Colocar el cursor en alguna sección del modelo de la proteína, al situar el cursor se marca en rojo el aminoácido del cuál se trata y su localización numérica.

7. Dar clic en la pestaña "*Templates*" en la cual aparece una lista de coincidencias entre la secuencia de la proteína introducida y las referencias guardadas en la base de datos *SWISS- PROT*.



Cuestionario y/o ejercicios complementarios

1. ¿Qué importancia tiene conocer la estructura tridimensional de las proteínas?

2. ¿Qué parámetro utiliza el servidor para modelar una proteína?

3. ¿Qué porcentaje de similitud existe entre la proteína más alejada reportada en las bases de datos con respecto a los datos obtenidos SWISS-MODEL?

4. Realizar el procedimiento con la proteína GFP y comparar los resultados obtenidos en ambas proteínas considerando la calidad del modelo, GMQE y QMEAN, discutan sus resultados.

Bibliografía

- Atwood, T., y Parry-Smith, D. (2002). *Introducción a la bioinformática*. Madrid, España: Pearson educación, S.A.

- Gómez, A., y Gómez, C. (2004). *Iniciación al estudio de la bioquímica*. Madrid, España: Grupo Anaya, S.A.

- Melo, V., y Cuamatzi, O. (2004). *Bioquímica de los procesos metabólicos*. Barcelona, España: Reverte, S.A.

- Oliva, R., y Vidal, M. (2006). *El genoma humano: nuevos avances en investigación, diagnóstico y tratamiento*. Barcelona, España: UBe.

- Peña, A., Arroyo, A., Gómez, A., y Tapia, R. (2018). *Bioquímica*. Ciudad de México, México: Limusa.

- Roldan, D. (2015). *Bioinformática, el ADN en un solo clic*. Bogotá, Colombia: Ediciones de la U.

- Waterhouse, A., Bertoni, M., Bienert, M.S., Studer, T., Gumienny, R., Heer, F.T., Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R. y Schwede, T. (2018). *SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acid Research.* 46(W1):W296-W303.

Práctica 10

Identificación de motivos conservados en secuencias de proteínas

MEME SUIT

Introducción

Los motivos o estructuras supersecundarias son combinaciones características de una estructura secundaria de 10 a 40 residuos de aminoácidos que se repiten en diferentes proteínas. Estos cierran la brecha entre la regularidad menos específica de una estructura secundaria y el plegamiento altamente específico de la estructura terciaria. Un mismo motivo puede encontrarse en diferentes proteínas con funciones similares, pueden actuar como unión a un ligando específico y contribuir a la estructura de un dominio. Algunos ejemplos de motivos son: proteínas hélice-giro-hélice, cremallera de leucina y dedos de zinc.

Herramienta *MEME*.

MEME (*Multiple Em for Motif Elicitation*) es una herramienta integrada por bases de datos que funcionan para identificar y analizar motivos en secuencias de aminoácidos en proteínas y nucleótidos en ácidos nucleicos usando modelos probabilísticos y discretos. *MEME* cuenta con múltiples opciones de búsqueda, entre las que se encuentra *MAST*, la cual busca en las bases de datos secuencias registradas que coincidan con cada motivo encontrado.

Objetivo general

- Identificación de motivos en una secuencia de aminoácidos de proteínas de diferentes organismos mediante una base de datos para determinar si los motivos encontrados son topológicos, estructurales o funcionales.

Objetivos particualres

1. Determinar con bases estadísticas la relación evolutiva de diversas especies para determinar qué tan cercana o alejada es una especie con respecto a otra.

2. Manejar las bases de datos MEME y obtener secuencias registradas que coincidan con cada motivo.

3. Comparar motivos de secuencias conocidas con la herramienta MAST con el fin de conocer motivos de otras especies registrados en las bases de datos.

Requerimientos para la práctica

- Dispositivo con conexión a Internet.

Nota: No es necesario descargar ningún programa para realizar esta práctica, se trabaja en línea.

Procedimiento

1. Durante esta práctica se realizará la identificación de motivos en un conjunto de secuencias de una proteína de diferentes especies y se determina la conservación de los motivos encontrados. En esta práctica se trabaja con las secuencias de la proteína *Max*. Obtener las secuencias en formato *FASTA* o los números de identificación (*Accession number*) de *NCBI/Protein*:

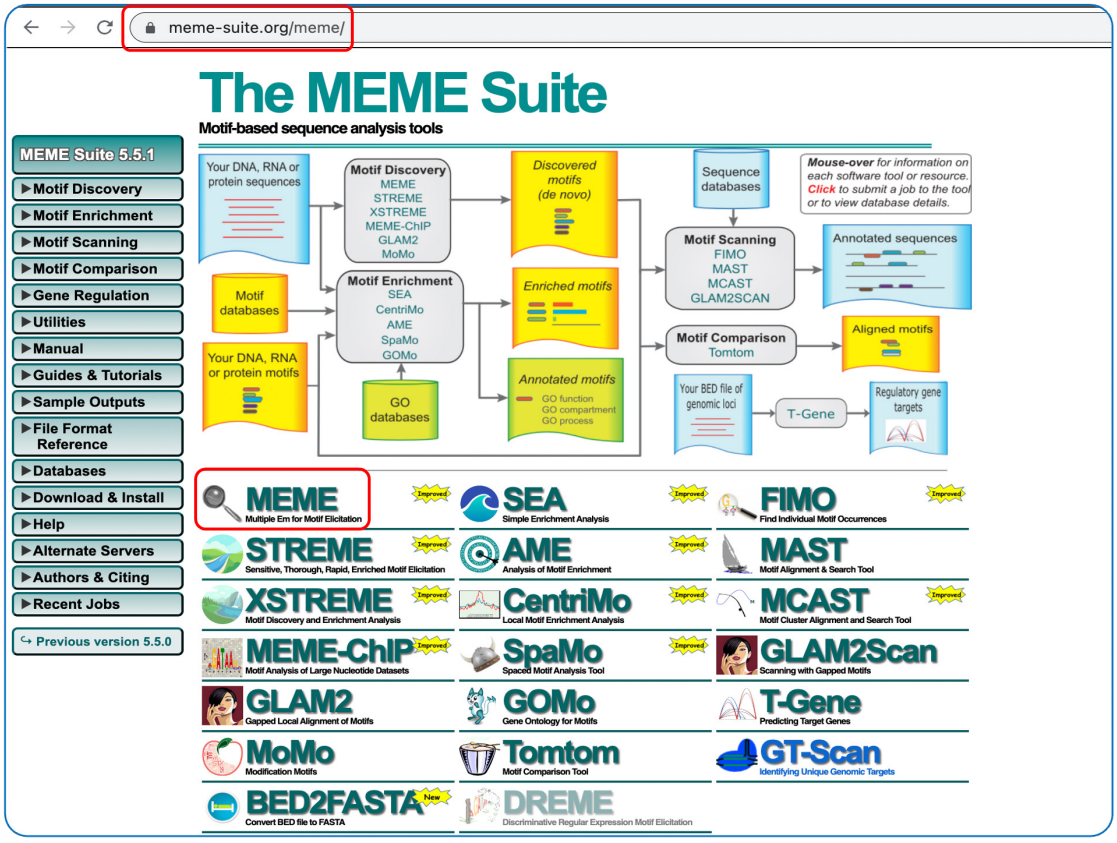
    - >AAI38673.1 *Max protein* [*Mus musculus*]

    - >AAH99778.1 *Max protein* [*Rattus norvegicus*]

- >RLQ65877.1 *Max* [*Cricetulus griseus*]
- >BAA07038.1 *Max* [*Felis catus*]
- >AAH81313.1 *Max protein* [*Xenopus tropicalis*]

2. En esta práctica se utilizarán las herramientas *MEME* (*Multiple Em for Motif Elicitation*) versión 5.4.1 *Suite* y *MAST*, ambas disponibles en https://meme-suite.org/meme/. Acceder a la plataforma *MEME*.

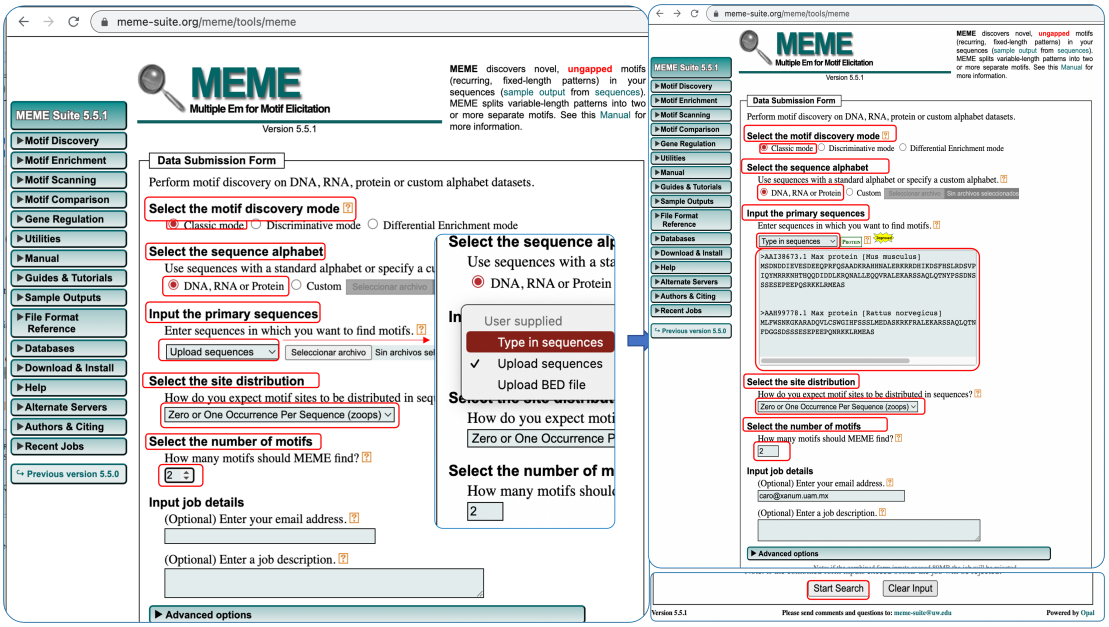
3. Una vez en la página principal de la suite de la herramienta *MEME*, situar el cursor sobre cada uno de los iconos que mostraran una breve descripción de la función de cada herramienta. Hacer clic en el icono correspondiente a *MEME*



4. Seleccionar los parámetros de búsqueda:

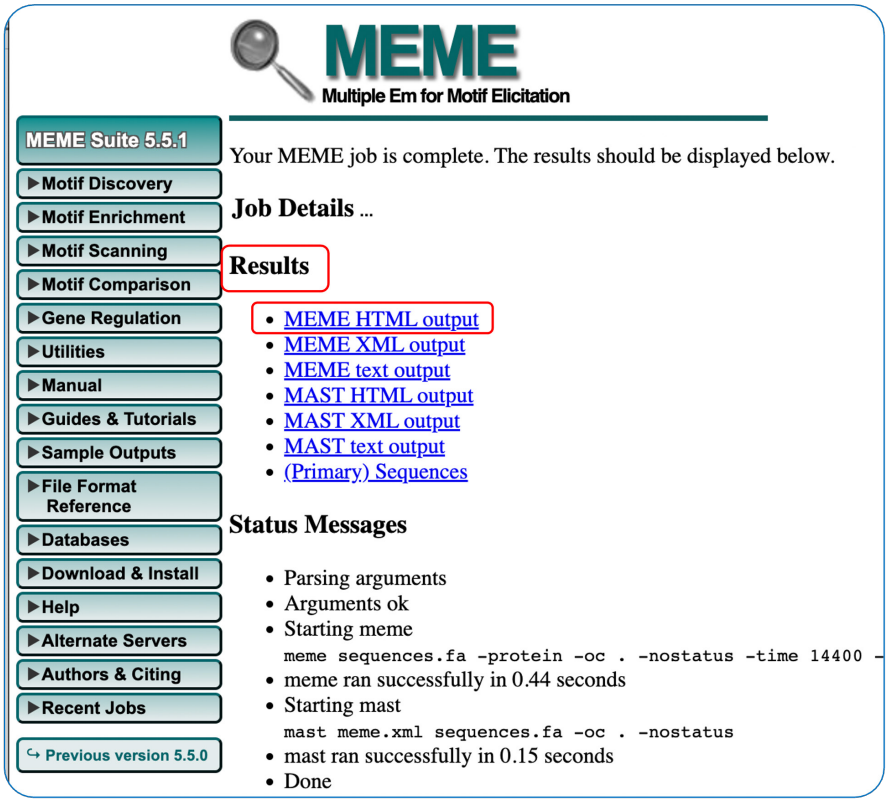
- "*Select the motif Discovery mode*" se elige la option "*Classic mode*"
- "*Select the sequence alphabet*" elegir "*DNA, RNA or Protein*"
- "*Input the primary sequences*" desplegar la pestaña y elegir "*Type in sequences*"
- "*Select the site distribution*", desplegar la pestaña para establecer el número de repeticiones que se espera encontrar de los motivos identificados a lo largo de cada secuencia. En este caso se indicará cada motivo que aparezca cero o una vez por secuencia, por lo que se elige la opción "*Zero or one occurrence per sequence*" (*zoops*).

- "*Select the number of motifs*", además, se establecerá el número de motivos que *MEME* debería encontrar. En esta sección se indican 2 motivos.



5. Utilizar los parámetros avanzados por defecto ("Advanced options") y presionar el botón "Start Search", en la parte inferior de la página. Dependiendo del número y longitud de las secuencias bajo estudio los resultados de MEME pueden tardar en aparecer. Una vez finalizado el análisis aparecen distintas opciones, se elige la primera de ellas, "MEME HTML output".

Esperar un momento



6. A continuación, se visualiza la sección "*Discovered Motifs*" donde se encuentran los 2 motivos. En este caso se observa cada motivo de 50 aminoácidos, de los cuales cada uno se ha identificado en 4 de las 5 secuencias analizadas, como indica la columna "*Sites*".

La representación gráfica indica las posiciones conservadas en el motivo, lo cual se mide en bits (para este propósito es suficiente saber que la altura de la columna es directamente proporcional a su nivel de conservación). Los colores corresponden al tipo de residuos con mayor prevalencia en dicha posición (por ejemplo, el color rojo hace referencia que en esa posición se encuentra principalmente aminoácidos cargados positivamente).

En esta sección se encuentra un resumen de la información sobre el motivo y su secuencia, que contiene la siguiente información:

- Valor E ("*E-value*"). Importancia estadística del motivo, *MEME* muestra los motivos de mayor importancia estadística (menor valor E). El cálculo del valor E de un motivo se basa en su rango de probabilidad, amplitud, sitios, frecuencia de las letras de fondo y el tamaño del conjunto de entrenamiento y constituye una estimación del número de motivos esperado dado un rango de probabilidad, y con la mínima amplitud y números de sitios que se encontraría de un conjunto de secuencias aleatorias de tamaño parecido.

- Amplitud ("*Width*"). Amplitud del motivo, cada motivo describe un patrón de una amplitud fija, puesto que *MEME* no permite huecos.

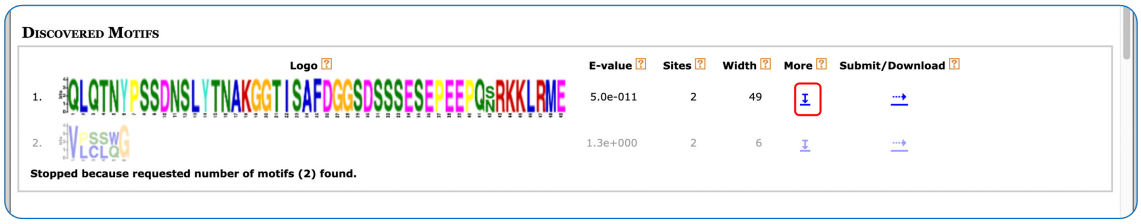
- Sitios ("*Sites*"). Número de sitios implicados en la construcción del motivo.

- Intervalo de probabilidad ("*Log likelihood ratio*"). Intervalo de probabilidad en unidades logarítmicas.

- Contenido de la información ("*Information content*"). Es el motivo de bits.

- Entropía relativa ("*Relative entropy*"). Es la entropía relativa del motivo en bits y se calcula como el intervalo de probabilidad dividido entre el número de sitios ("*Sites*").

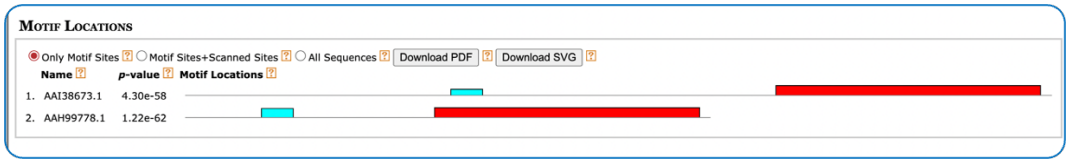
7. Dar clic para obtener la información específica en "*More*" de cada motivo.



8. A continuación, se muestra la información sobre los sitios del motivo dados por *MEME*, representados por colores y diez posiciones anteriores y posteriores que se muestran en gris que no forman parte del motivo.

9. Se puede descargar una imagen en alta calidad de cada uno de los motivos identificados haciendo clic en la flecha de la columna "*Submit/Download*". En la pestaña "*Download logo*" se elige el formato "*PNG*" y finalmente se pulsa el botón "*Download*". Como se muestra en el siguiente ejemplo:

10. En la sección "*Motif Locations*", se muestran los motivos a lo largo de cada una de las secuencias comparadas en el diagrama combinado de bloques. Se sugiere descargar en formato PDF para obtener una imagen en la que observe de manera definida la secuencia de cada motivo mostrada gráficamente.



11. En la siguiente sección "*Inputs & Settings*", se muestra un resumen con los detalles de los datos que *MEME* utilizó para identificar los motivos; como el número de secuencias problema y la frecuencia de cada residuo aminoacídico, de este conjunto de secuencias y además los parámetros utilizados para realizar el análisis.

Inputs & Settings

Sequences

Role	Source	Alphabet	Sequence Count	Total Size
Primary Sequences	sequences.fa	Protein	2	257

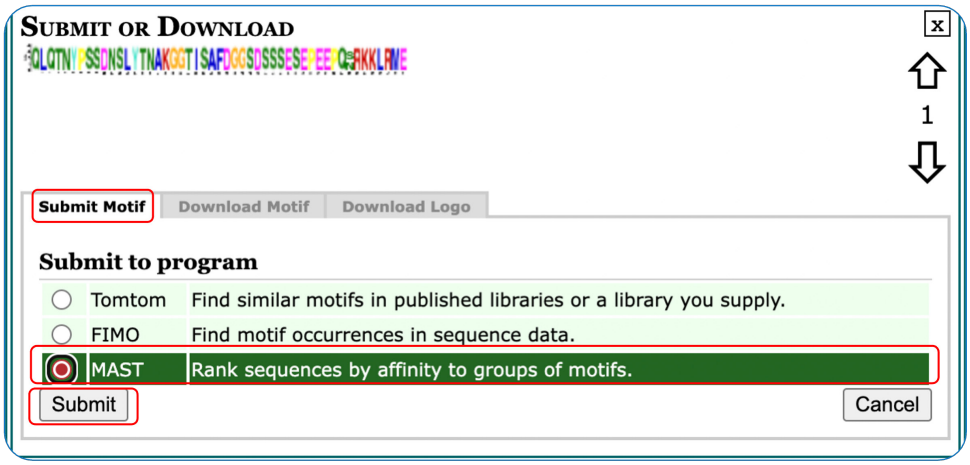
Background Model

Source: built from the (primary) sequences

Order: 0

	Name	Freq.	Bg.
A	Alanine	0.0895	0.0895
C	Cysteine	0.00389	0.00391
D	Aspartic acid	0.0778	0.0778
E	Glutamic acid	0.0817	0.0817
F	Phenylalanine	0.0272	0.0272
G	Glycine	0.0428	0.0428
H	Histidine	0.0272	0.0272
I	Isoleucine	0.0311	0.0311
K	Lysine	0.0739	0.0739
L	Leucine	0.07	0.07
M	Methionine	0.0233	0.0234
N	Asparagine	0.0467	0.0467
P	Proline	0.0311	0.0311
Q	Glutamine	0.0661	0.0661
R	Arginine	0.0778	0.0778
S	Serine	0.152	0.152
T	Threonine	0.0311	0.0311
V	Valine	0.0156	0.0156
W	Tryptophan	0.00778	0.0078
Y	Tyrosine	0.0233	0.0234

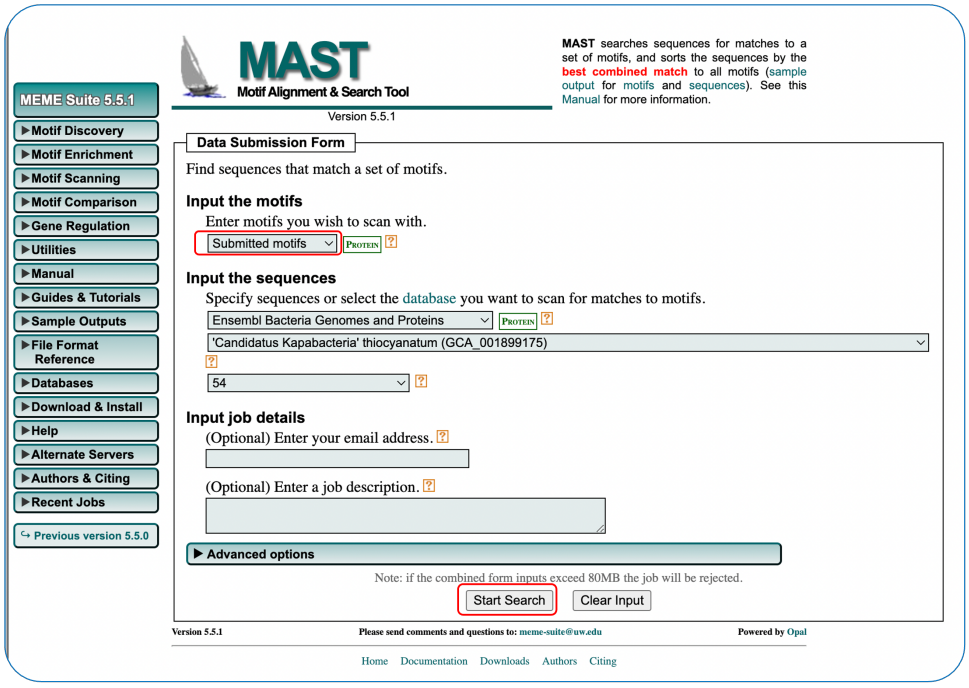
12. Una vez que se han identificado los posibles motivos, se puede comprobar si están presentes en otras secuencias. Esto se consigue gracias a la herramienta de la *Suite MEME* llamada *MAST*, para buscar coincidencias de cada motivo, ir a la sección "*Discovered motifs*", dar clic en "*Submit/Download*" de cada motivo, que muestra una ventana, en la cual se elige la pestaña "*Submit Motif*", seleccionar la opción *MAST*. Dar clic en "*Submit*".



13. De la pantalla que se muestra a continuación, realizar la búsqueda eligiendo los siguientes criterios para seleccionar la base de datos.

    En "*Input the Sequences*" seleccionar:

    - Coincidencias de un conjunto de genomas y proteínas.
    - La especie en la que se buscarán las coincidencias: Puede elegir una especie cercana y otra alejada evolutivamente.
    - La versión disponible de la base de datos: elegir la última versión.
    - Dar clic en "*Start Search*".



14. Los resultados de *MAST* pueden tardar en aparecer. Una vez finalizado el análisis, hacer clic en "*MAST HTML output*"

Esperar un momento

MAST
Motif Alignment & Search Tool

MEME Suite 5.5.1
▶Motif Discovery
▶Motif Enrichment
▶Motif Scanning
▶Motif Comparison
▶Gene Regulation
▶Utilities
▶Manual
▶Guides & Tutorials
▶Sample Outputs
▶File Format Reference
▶Databases
▶Download & Install
▶Help
▶Alternate Servers
▶Authors & Citing
▶Recent Jobs
↪ Previous version 5.5.0

Your MAST job is complete. The results should be displayed below.

Job Details ...

Results

- MAST HTML output
- MAST XML output
- MAST text output
- Input Motifs

Status Messages

- Parsing arguments
- Arguments ok
- Starting mast

```
mast -oc . -nostatus -bfile db/EnsemblBacteria/_candidatus_
10.0 -df EnsemblBacteria/_candidatus_kapabacteria_thiocyana
query=SEQUENCEID&sort=score motifs.meme db/EnsemblBacteria/
```

- mast ran successfully in 0.24 seconds
- Done

15. Las secuencias encontradas se ordenan por el valor de *E* y *p* de la secuencia, de menor a mayor. Seguido de la secuencia y del valor *E* aparece una flecha que al dar clic despliega información adicional de la secuencia en un panel que se abrirá. A continuación, se muestra el diagrama de bloques de motivos que se han encontrado de las secuencias en las bases de datos. En cuanto a la información adicional, se muestra la descripción de la secuencia, el Valor *p*, combinado y la secuencia anotada indicando la posición del motivo dentro de la misma.

Cuestionario y/o ejercicios complementarios

1. Investigar qué parte del motivo interacciona con el DNA.
2. ¿Cómo podrías calcular que tan alejados o cercanos evolutivamente se encuentran los motivos obtenidos por MAST (Valor p/Valor E)?
3. Investigar algunos ejemplos de motivos y su función.

Bibliografía

- Atwood, T., y Parry-Smith, D. (2002). *Introducción a la bioinformática*. Madrid, España: Pearson educación, S.A.
- Capel, J., y Yuste, F. (2016). *Manual de prácticas de bioinformática*. Andalucía España: Almería.
- Luque, J. (2001). *Biología molecular e ingeniería genética*. Madrid, España: Harcourt.
- Pelley, J. (2011). *Elsevier's Integrated Review Biochemistry*. 2nd Edition. Saunders. EUA.
- Oliva, R., y Vidal, J. (2006). *El genoma humano: nuevos avances en investigación, diagnóstico y tratamiento*. Barcelona, España: UBe.
- Roldan, D. (2015). *Bioinformática, el ADN en un solo clic. Bogotá*, Colombia: Ediciones de la U.
- Stryer, L., Berg, J., y Tymokzko, J. (2013). *Bioquímica*. Madrid, España: Reverté, S.A.

Práctica 11

Predicción de la ubicación subcelular de las proteínas

PSORT/PSORTII

Introducción

La mayoría de las proteínas son sintetizadas en el retículo endoplásmico rugoso mediante los ribosomas asociados a su membrana y otras en el citosol por los ribosomas. Una minoría por los cloroplastos y las mitocondrias.

Por su diversidad, las proteínas se han clasificado por su jerarquía, composición, conformación, solubilidad, función, estructura y localización celular. La relación estructura-forma-función está íntimamente relacionada a la ubicación de las proteínas. De acuerdo con la posición en la célula se pueden localizar:

Citosólicas

- Las proteínas "solubles" no están localizadas en ningún organelo en particular. Son componentes del citosol, entre las que se encuentran las enzimas que funcionan como centros catalíticos individuales, que actúan sobre metabolitos que están en disolución en el citosol.

- Estructuras macromoleculares formadas por proteínas (y a veces otros componentes) pueden estar localizadas en sitios concretos del citoplasma; por ejemplo, los centriolos que están asociados a las regiones polares. Ejemplo las proteínas involucradas en la glucólisis.

Nucleares

- Las proteínas nucleares han de ser transportadas desde el citosol donde se sintetizan hacia el núcleo mismas que atraviesan la membrana nuclear, antes de que puedan ocupar su sitio en el núcleo. Muchas proteínas nucleares son componentes de la cromatina propiamente dicha; otras forman parte de la lámina nuclear o de la matriz. Ejemplos en la membrana nuclear las proteínas del poro y al interior las ribonucleoproteínas.

Organelos

- Los organelos contienen proteínas sintetizadas en el citosol que son transportadas específicamente a la (o a través de) membrana del organelo al que están destinadas. Ejemplos en mitocondrias, en la membrana interna los cinco complejos proteicos, en la membrana del lisosoma la V-ATPasa y al interior diversas enzimas como proteasas.

Secreción

- Las proteínas secretadas por la célula llegan al exterior pasando a través de la membrana plasmática. Las síntesis de estas proteínas comienzan de la misma forma que de las proteínas asociadas a la membrana, pero recorren enteramente el sistema de transporte, en lugar de quedarse en algún punto intermedio del recorrido. Ejemplos: colágeno, elastina y albúmina

Objetivo general

- Predecir la ubicación de las proteínas mediante *PSORT* con el fin de relacionarlo con la función.

Objetivos particulares

1. Interpretar la información de la ubicación del proteoma para predecir la relación estructura, función y localización en la célula.

2. Predecir el destino final de las proteínas mediante *PSORT* con el fin de determinar la ubicación y función extracelular biológica.

Requerimientos para la práctica

- Dispositivo con conexión a Internet.

Nota: No es necesario descargar ningún programa para realizar esta práctica, se trabaja en línea.

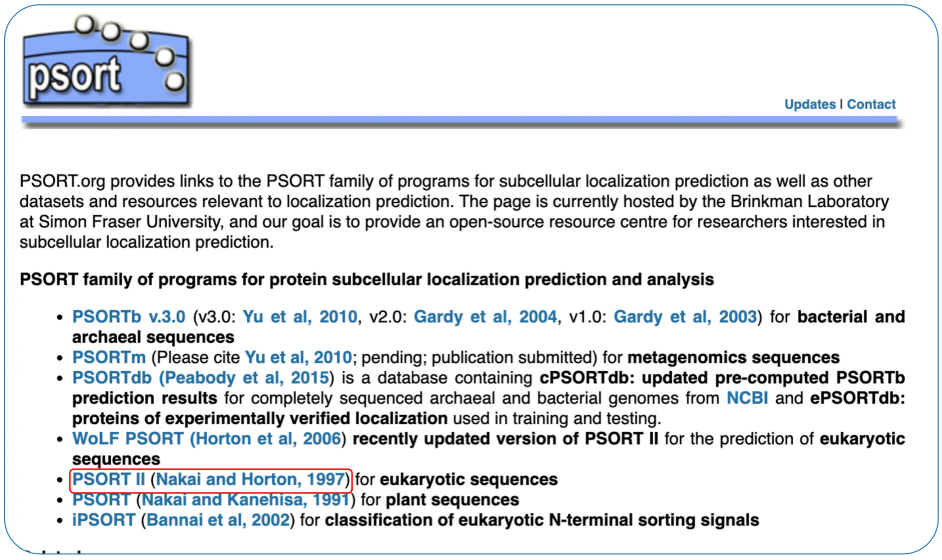
Procedimiento

Esta práctica se realizará con las siguientes proteínas:

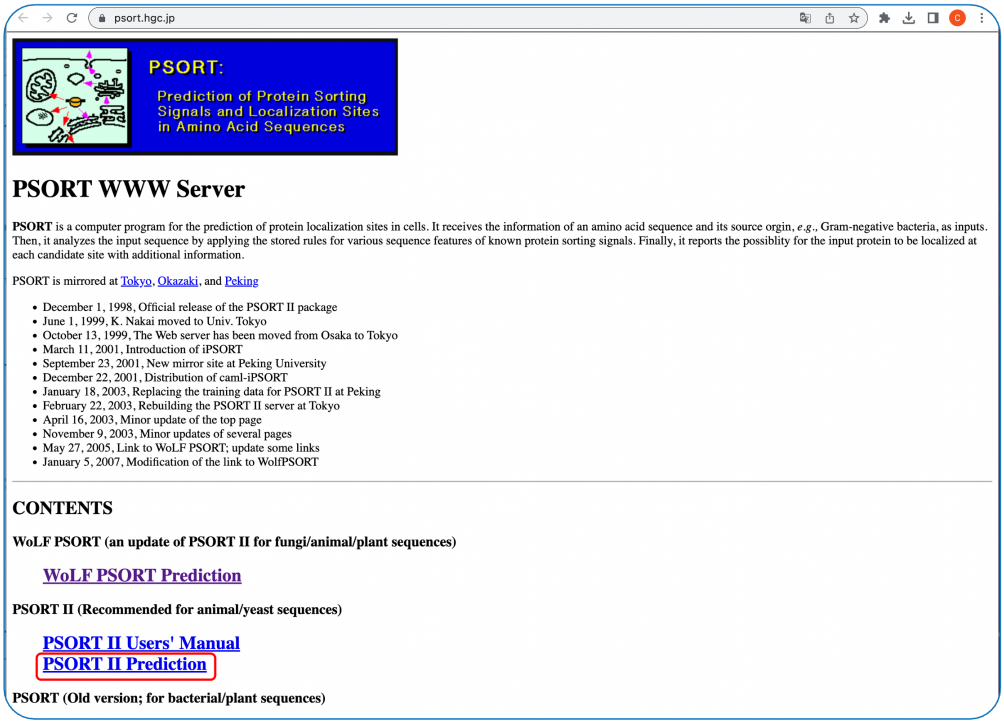
- *Insulin* [*Homo sapiens*].
- *Green-fluorescent protein* [*Aequorea victoria*].
- *Class II beta tubulin isotype* [*Homo sapiens*].
- *Hepatocyte nuclear factor 1-alpha isoform 2* [*Homo sapiens*].
- *Cation-transporting ATPase* 13A2 *isoform* 1 [*Homo sapiens*].

1. Obtener el formato *FASTA* de las proteínas en *NCBI/Protein/FASTA*.

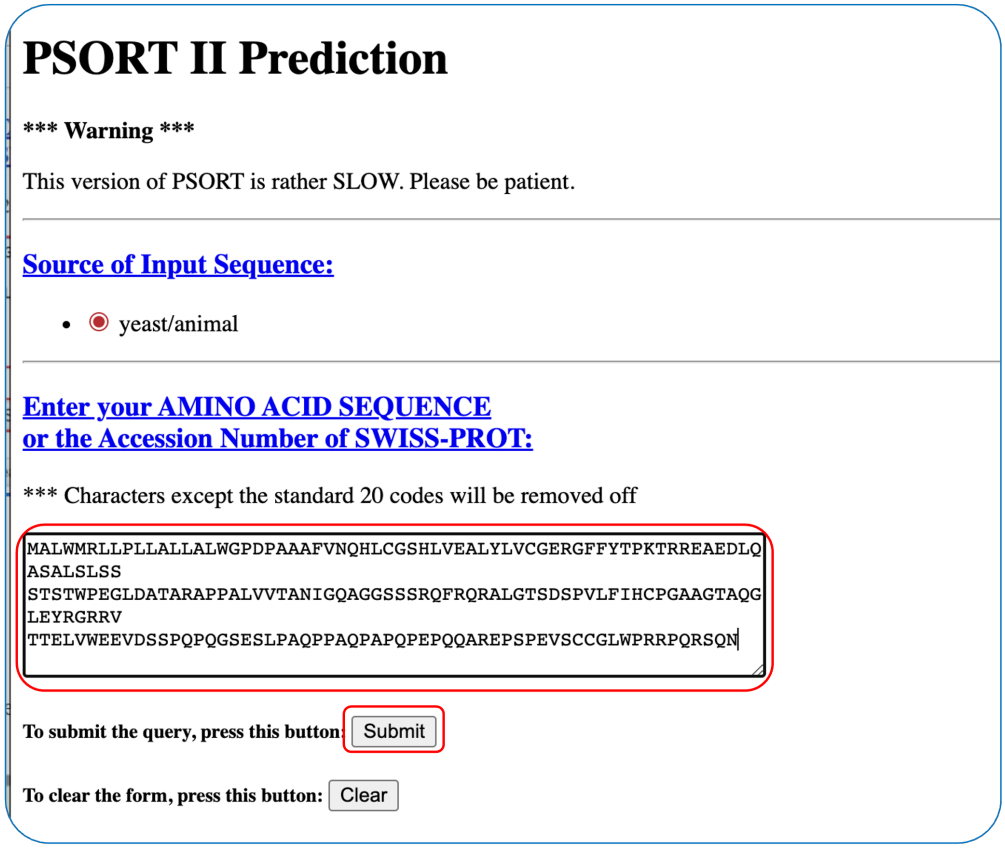
2. Acceder a la página principal de *PSORT* en el siguiente enlace https://www.psort.org/. En esta práctica se utilizan las herramientas *Wolf PSORT* y *PSORT* II. En el caso de que se requiera predecir la ubicación celular de una proteína en plantas, se utiliza la herramienta *PSORT*. Adicionalmente se encuentran recursos relacionados con la predicción de la ubicación subcelular de las proteínas.



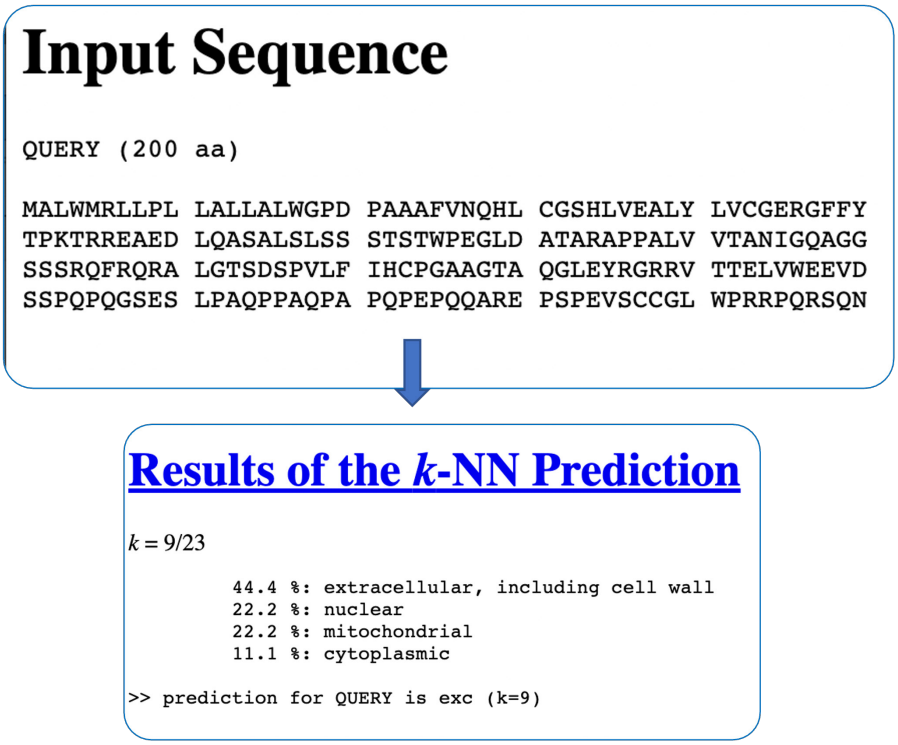
3. Dar clic en *PSORT* II, se abrirá la siguiente pantalla con una breve introducción del servidor.



4. Dar clic en "*PSORT* II *Prediction*" e introducir la secuencia de la proteína en formato *FASTA*. Dar clic en "*Submit*". El procedimiento se realiza por separado con cada una de las proteínas, únicamente se muestran los recuadros del proceso para obtener la localización de la insulina.



Aparece una página de resultados que muestra la secuencia introducida seguida de los resultados de los subprogramas "*Results of Subprograms*" que indica cada método de predicción utilizado por el servidor *PSORT*. La parte final nos muestra los resultados de la predicción subcelular de la proteína, realizado mediante el algoritmo k-NN, indicando el porcentaje de probabilidad de la localización de la proteína en la célula o extracelular.



Input Sequence

QUERY (200 aa)

MALWMRLLPL LALLALWGPD PAAAFVNQHL CGSHLVEALY LVCGERGFFY
TPKTRREAED LQASALSLSS STSTWPEGLD ATARAPPALV VTANIGQAGG
SSSRQFRQRA LGTSDSPVLF IHCPGAAGTA QGLEYRGRRV TTELVWEEVD
SSPQPQGSES LPAQPPAQPA PQPEPQQARE PSPEVSCCGL WPRRPQRSQN

Results of the *k*-NN Prediction

$k = 9/23$

44.4 %: extracellular, including cell wall
22.2 %: nuclear
22.2 %: mitochondrial
11.1 %: cytoplasmic

>> prediction for QUERY is exc (k=9)

En la siguiente tabla se muestra el resultado de la predicción k-NN de la sublocalización celular de 6 proteínas:

Insulin [*Homo sapiens*] 66.7 %: extracellular, including cell wall 11.1 %: endoplasmic reticulum 11.1 %: mitochondrial 11.1 %: vacuolar	**Green-fluorescent protein** [*Aequorea victoria*] 60.9 %: cytoplasmic 17.4 %: nuclear 4.3 %: vesicles of secretory system 4.3 %: cytoskeletal 4.3 %: mitochondrial 4.3 %: Golgi 4.3 %: vacuolar
RecName: Full=Collagen alpha-1(I) chain; AltName: Full=Alpha-1 type I collagen; Flags: Precursor 55.6 %: extracellular, including cell wall 22.2 %: cytoplasmic 11.1 %: nuclear 11.1 %: vacuolar	**Class II beta tubulin isotype** [*Homo sapiens*] 55.6 %: cytoskeletal 44.4 %: cytoplasmic
Hepatocyte nuclear factor 1-alpha isoform 2 [*Homo sapiens*] 69.6 %: nuclear 13.0 %: cytoplasmic 8.7 %: cytoskeletal 4.3 %: mitochondrial 4.3 %: Golgi	**Cation-transporting ATPase 13A2 isoform 1** [*Homo sapiens*] 69.6 %: plasma membrane 13.0 %: endoplasmic reticulum 8.7 %: vacuolar 4.3 %: nuclear 4.3 %: Golgi

Cuestionario y/o ejercicios complementarios

1. Relacione la estructura y función de las proteínas del cuadro anterior con la localización.
2. Realice el ensayo con *PSORT/PSORTII* de 4 proteínas con (membrana plasmática, membrana mitocondrial, citosólica y secreción).
3. Analice los porcentajes de la localización subcelular de los resultados de las proteínas anteriores.

Bibliografía

- Atwood, T., y Parry-Smith, D. (2002). *Introducción a la bioinformática*. Madrid, España: Pearson educación, S.A.
- Capel, J., y Yuste, F. (2016). *Manual de prácticas de bioinformática*. Andalucía España: Almería.
- Lewin, B. (1996). *Genes*. Barcelona España: Reverté
- Roldan, D. (2015). *Bioinformática, el ADN en un solo clic*. Bogotá, Colombia: Ediciones de la U.

Práctica 12

Acoplamiento molecular (*docking*)

PyRx

Introducción

El acoplamiento molecular (*molecular docking*) es un método computacional que permite predecir sitios de unión entre moléculas candidatas a fármacos (ligando) y un receptor (diana o blanco). Es importante mencionar que se pueden establecer acoplamientos entre proteína-proteína, proteína- carbohidrato, antígeno-anticuerpo y actualmente los estudios de *molecular docking* juegan un papel muy importante en la interacción de biomaterial-ligando o proteína-biomaterial, etc. Tanto el receptor como el ligando pueden actuar como una proteína, carbohidrato, anticuerpo, antígeno y/o biomaterial.

Existe una gran variedad de programas donde se pueden llevar a cabo estudios de acoplamiento molecular y cada programa tiene un algoritmo y una función de *scoring* diferente. Esta función define las poses correctas de las incorrectas a través de un criterio energético, es decir, calculando las afinidades de unión entre el receptor y el ligando.

Asimismo, se han realizado numerosos protocolos en el acoplamiento receptor-ligando de tipo rígido o flexible dependiendo de las simulaciones, por ejemplo:

a) Acoplamiento de ligando flexible y mantener rígido al receptor.

b) Acoplamiento rígido (receptor y ligando rígidos).

c) Acoplamiento flexible (ambas moléculas flexibles).

Los programas más utilizados para estos estudios *in silico* son: *AutoDock Vina*, *PyMOL*, *PyRx, entre otros*.

Es importante mencionar que estos programas utilizan las coordenadas de una caja para los cálculos de acoplamiento entre el receptor y el ligando, es decir, si se conoce el sitio de unión de interés, se reducen las dimensiones de la caja en la zona correspondiente. En caso de no conocer el sitio de unión se sugiere que el tamaño de la caja abarque la superficie completa del receptor.

En esta práctica se utilizará el programa *PyRx* para realizar los estudios de acoplamiento molecular. También, se recomienda usar el programa *PyRx* para aprender está técnica de acoplamiento molecular, ya que es de licencia libre y la interfaz es muy amigable. En el programa de *PyRx*, el receptor de interés se denominará la macromolécula y el fármaco como ligando. Después de definir estos dos, se ejecutarán a través de *AutoDock* Vina de manera predeterminada para obtener resultados confiables, los cuales se pueden procesar para obtener más información y poder contribuir a la diversidad de resultados en investigación.

Objetivo general

- Realizar estudios de acoplamiento molecular entre una proteína y un ligando de interés a través del programa *PyRx 0.8*.

Objetivos particulares

1. Analizar la mayoría de las poses o confórmeros obtenidos para tener un mejor muestreo de la interacción proteína-ligando.

2. Comprobar si alguno de los sitios encontrados por el programa *PyRx 0.8* concuerda con el sitio de unión reportado en la literatura del sistema bajo estudio.

Requerimientos para la práctica

- Dispositivo con conexión a Internet.

- Descargar el programa *PyRx 0.8* en la página https://pyrx.sourceforge.io/ dependerá del sistema operativo que requiera el alumno *Windows*, *MacOs* o *Linux*. Otra opción, es descargar el ejecutable de *PyRx 0.8* para *Windows* o *MacOS* https://drive.google.com/drive/u/1/folders/1CLu30e0w9OZA_Ys-ccZKmKfYaBifnfRv

- Descargar el programa *PyMOL* en la página https://pymol.org/2/ dependerá del sistema operativo que requiera el alumno *Windows*, *MacOs* o *Linux*. Otra opción, es descargar el ejecutable de *PyMOL* para *Windows*, *MacOs* o *Linux* https://drive.google.com/drive/u/1/folders/1yaGAbjQpV9TET8JNsB1LK0mla0A-kSMM

- Descargar los archivos pdb para llevar a cabo el acoplamiento molecular, en esta práctica se utilizará como ejemplo el receptor para productos finales de glucosilación avanzada (*rage*) y como ligando el ácido quinolínico (*quin*) https://drive.google.com/drive/folders/1X1QCU-YqjwaJhOHDPnZWyvh_n-fdu1yS?usp=sharing

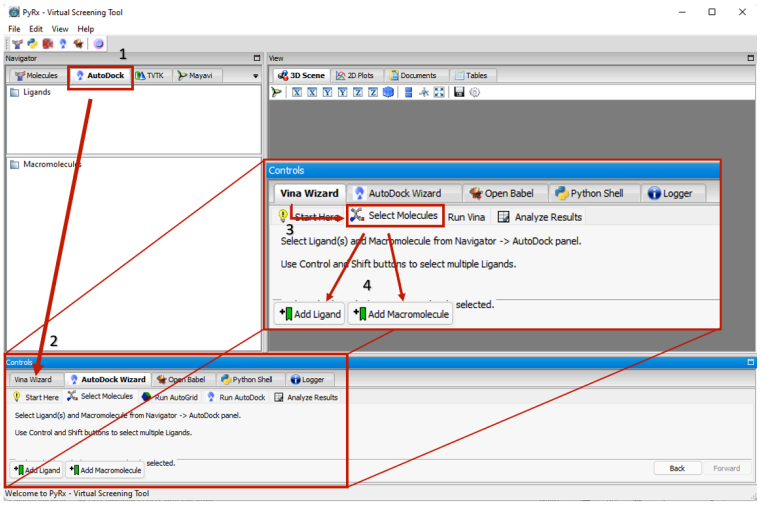
Procedimiento

1. Para obtener el archivo de la proteína de interés con extensión .pdb se busca en el *Protein Data Bank* (*PDB*) https://www.rcsb.org/, o si ya se cuenta con un archivo que contenga los datos estructurales de una proteína se puede utilizar y también debe tener extensión .pdb para poder visualizarlo en el programa *PyRx 0.8*. Para esta práctica se utilizará el receptor RAGE el cual se buscará en la página del PDB con el código 3CJJ.pdb y se descargará el archivo (3CJJ.pdb). Cabe mencionar que, los archivos .pdb pueden abrirse con *word* y se pueden borrar ligandos o moléculas de agua que contenga dicho archivo para que no intervengan en los estudios de acoplamiento molecular. La otra opción es descargar el archivo rage.pdb que está en la liga que se mencionó en el apartado anterior (**Requerimientos para la práctica**)

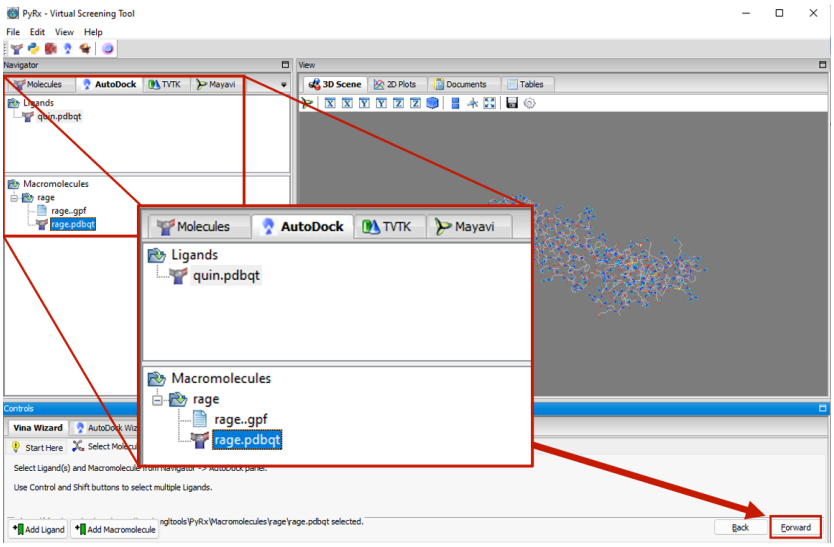
2. El ligando o fármaco de interés también debe tener extensión .pdb, para facilitar el desarrollo de la práctica se puede descargar de la liga mencionada en el apartado de **Requerimientos para la práctica**. Es importante mencionar que en caso de no contar con la estructura del ligando que se va a utilizar, se podrá construir con algún programa, por ejemplo, Avogadro, el cual es de licencia libre y de interfaz amigable e intuitiva.

   El programa Avogadro se puede descargar mediante la liga https://sourceforge.net/projects/avogadro/files/latest/download.

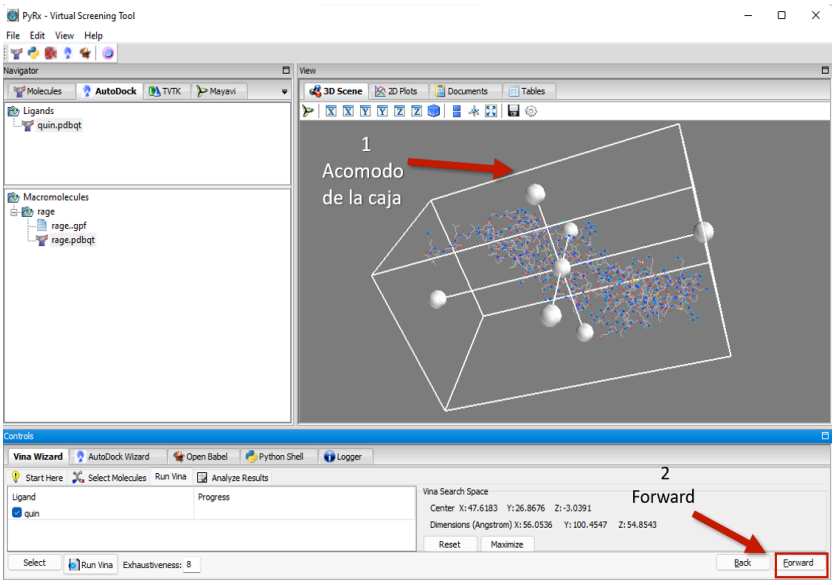
3. Una vez listos ambos archivos en extensión .pdb (proteína y ligando), se cargarán en el programa *PyRx 0.8*. Al abrir dicho programa se tiene la siguiente pantalla y se trabajará en la ventana de "*AutoDock Vina*". Posteriormente, se puede adicionar la proteína con el botón que dice "*Add Macromolecule*" y el ligando con "*Add Ligand*" como se muestra a continuación.



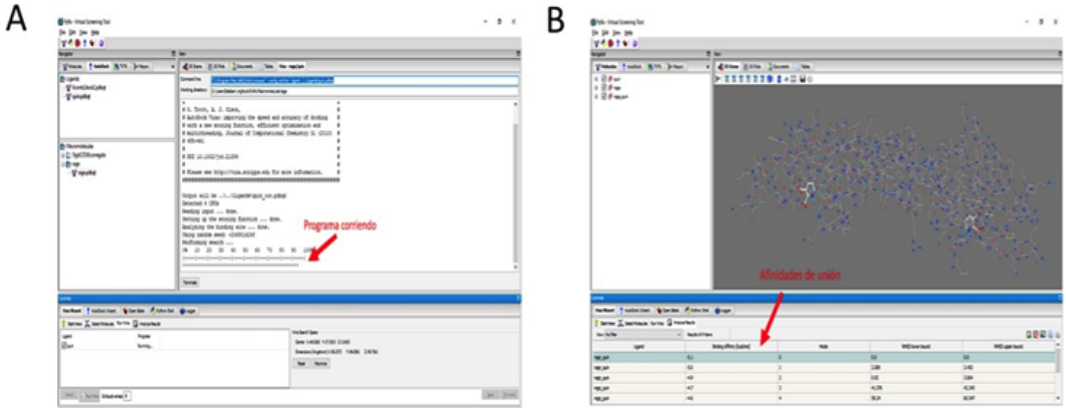
4. Una vez adicionadas ambas moléculas se observará que en las ventanas de lado izquierdo aparecerá en la parte superior una carpeta que dice "*Ligands*" y otra "*Macromolecules*". Respectivamente, en cada carpeta se registra el archivo que se va adicionando y se cambia la extensión de los archivos .pdb en. pdbqt. Esta extensión (.pdbqt) indica que se adicionaron cargas a cada átomo tanto en el archivo de la proteína como en el archivo del ligando. Posteriormente, ya cargadas ambas moléculas se presiona el botón "*Forward*". Cabe mencionar que, al presionar el botón "*Forward*" en ocasiones es necesario cargar nuevamente la Macromolécula o Ligando eso lo indicará el programa.



5. Al presionar "*Forward*" aparece una caja, la cual permite seleccionar la región del sitio de unión (en caso de que ya esté reportado en la literatura). En caso de no conocer este sitio, se sugiere que la caja cubra toda la superficie de la proteína. Una vez establecidas las dimensiones de la caja, se presionará nuevamente el botón "*Forward*".

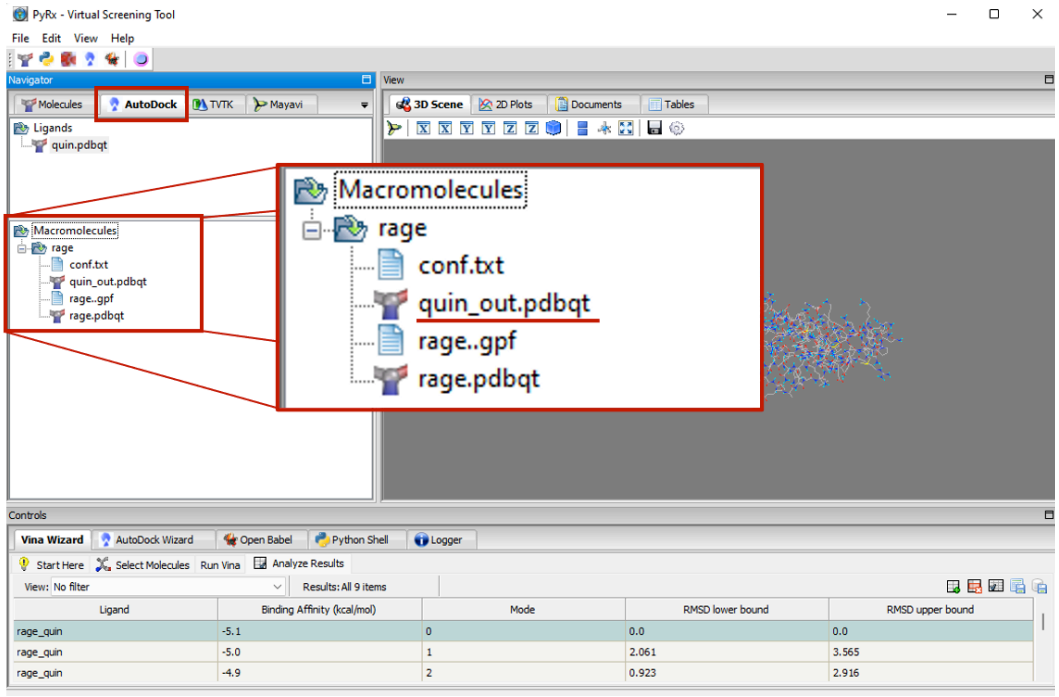


6. Una vez que hayan corrido los cálculos de acoplamiento molecular (panel A), estos resultados se observarán en una Tabla, donde se especificarán los valores de las afinidades de unión (kcal/mol) de cada una de las poses o confórmeros generados (panel B).

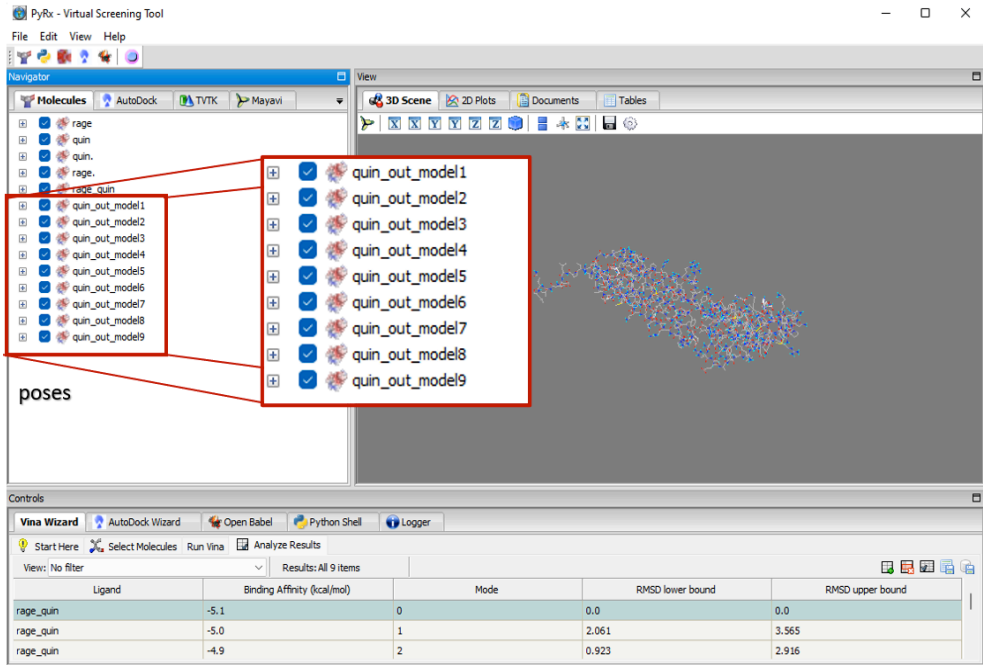


7. Existen diferentes formas para extraer las poses en formato .pdb:

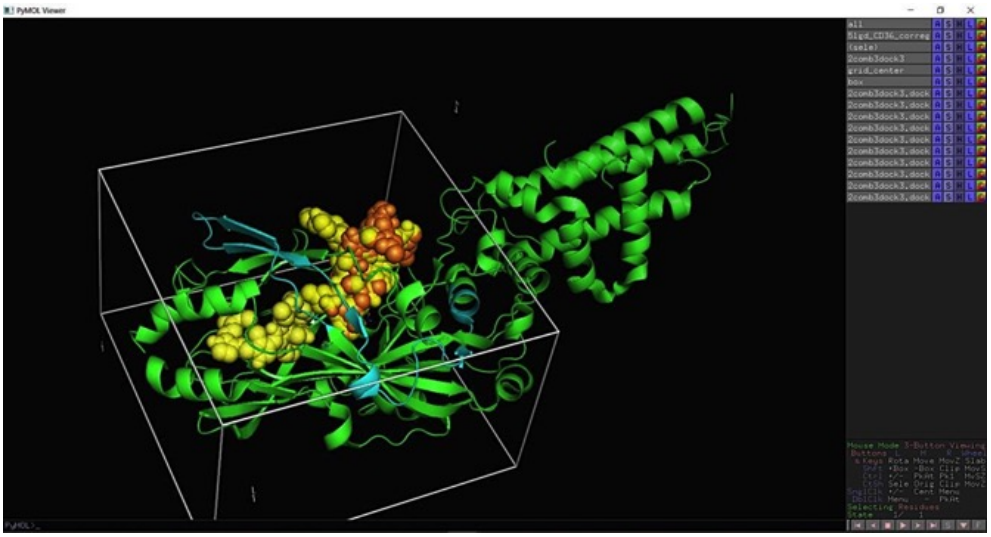
a) Una de ellas es dar clic en la ventana de "*AutoDock*" y aparecerá la carpeta de "*Macromolecule*", la cual contiene tres archivos adicionales: *protein.pdbqt* (*rage.pdbqt*), *ligand_out.pdbqt* (*quin_out.pdbqt*) y *conf.txt*. Sí se da clic en *ligand_out.pdbqt* (*quin_out.pdbqt*), inmediatamente se desplegarán los archivos de las poses de forma muy rápida.



Para trabajar con cada archivo se dará clic en la ventana que dice "*Molecules*" y se desplegará la lista de las poses obtenidas. Para guardar cada archivo en formato .pdb sólo se dará clic derecho y aparecerá "*Save PDB*", tanto la proteína como cada una de las poses pueden ser visualizados en el programa *PyMOL*. Recuerde que ambos archivos deben tener formato .pdb.



b) La otra opción para visualizar las poses obtenidas es exportar el archivo de la proteína.pdb (*rage.pdb*) y el archivo *ligand_out.pdbqt* (*quin_out.pdbqt*) al programa *PyMOL*. En la siguiente práctica se mostrará cómo exportar estos resultados.



Bibliografía

- Chapter 19. Dallakyan, S. y Olson, A. J. (2014). Small-molecule library screening by docking with PyRx. *Chemical Biology*. 243-250.
- *The PyMOL Molecular Graphics System*, Version 1.2r3pre, Schödinger, LLC.
- Trott, O. and Olson, A. J. A. (2010). *AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading*. Journal of Computational Chemistry 31: 455-461.
- *Avogadro: an open-source molecular builder and visualization tool*. Version 1.XX. http://avogadro.cc/
- Hanwell, M.D., Curtis, D.E., Lonie, D.C., Vandermeersch, T., Zurek, E. and Hutchison, G.R (2012). *Avogadro: An advanced semantic chemical editor, visualization, and analysis platform*. Journal of Cheminformatics. 4:17.

Práctica 13.

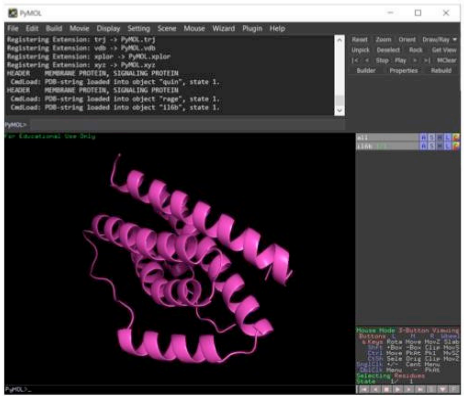
Visualización de los complejos obtenidos por los estudios de acoplamiento molecular

PyMOL

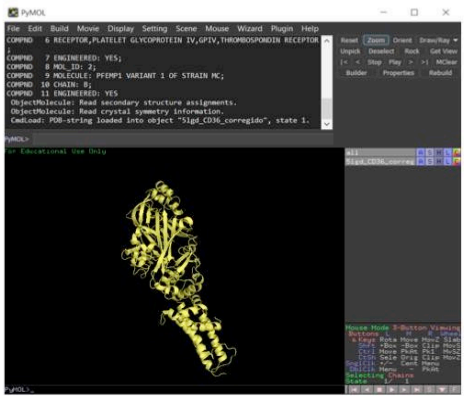
Introducción

Actualmente existe una gran cantidad de programas que presentan una interfaz de usuario gráfica para visualizar los archivos de coordenadas atómicas, uno de ellos es *PyMOL*. Este programa es ampliamente utilizado en biología estructural, ya que permite producir imágenes tridimensionales de alta calidad de proteínas, moléculas pequeñas e incluso nanomateriales de carbono como grafeno, nanotubos de carbono, fullerenos, etc. *PyMOL* es un software ampliamente usado y difundido en el ámbito científico, ya que cuenta con una licencia libre para alumnos e investigadores. La utilidad de *PyMOL* es muy amplia, sin embargo, su principal aplicación es la visualización de una biomolécula en 3D como se muestra en la siguiente Figura.

A



B



En la práctica anterior se llevaron a cabo estudios de acoplamiento molecular sobre un sistema en cuestión y estos resultados se pueden visualizar con el programa *PyMOL*. Este programa también permitirá guardar cada complejo en formato .pdb, lo cual es muy importante ya que se necesita el archivo complejo.pdb para llevar a cabo otro tipo de estudios en diferentes programas para estudiar exhaustivamente el reconocimiento molecular proteína-ligando, por ejemplo, generando mapas de interacción, realizar estudios de dinámica molecular, estudios de cálculos de energía de unión, entre otros.

Objetivo general

- Visualizar los confórmeros obtenidos por el estudio de acoplamiento molecular realizado en la práctica anterior a través del programa *PyMOL*.

Objetivos particulares

1. Analizar las diferentes conformaciones de los ligandos en la proteína encontrados a partir del estudio de acoplamiento molecular de proteína-ligando.

2. Validar si el sitio o sitios encontrados concuerdan con el reportado en la literatura del sistema en cuestión.

Requerimientos para la práctica

- Computadora con conexión a Internet.

- Descargar el programa *PyRx 0.8* en la página https://pyrx.sourceforge.io/ dependerá del sistema operativo que requiera el alumno *Windows*, *MacOs* o *Linux*. Otra opción, es descargar el ejecutable de *PyRx 0.8* para *Windows* o *MacOS* https://drive.google.com/drive/u/1/folders/1CLu30e0w9OZA_Ys-ccZKmKfYaBifnfRv

- Descargar el programa *PyMOL* en la página https://pymol.org/2/ dependerá del sistema operativo que requiera el alumno *Windows*, *MacOs* o *Linux*. Otra opción, es descargar el ejecutable de *PyMOL* para *Windows*, *MacOs* o *Linux* https://drive.google.com/drive/u/1/folders/1yaGAbjQpV9TET8JNsB1LK0mla0A-kSMM

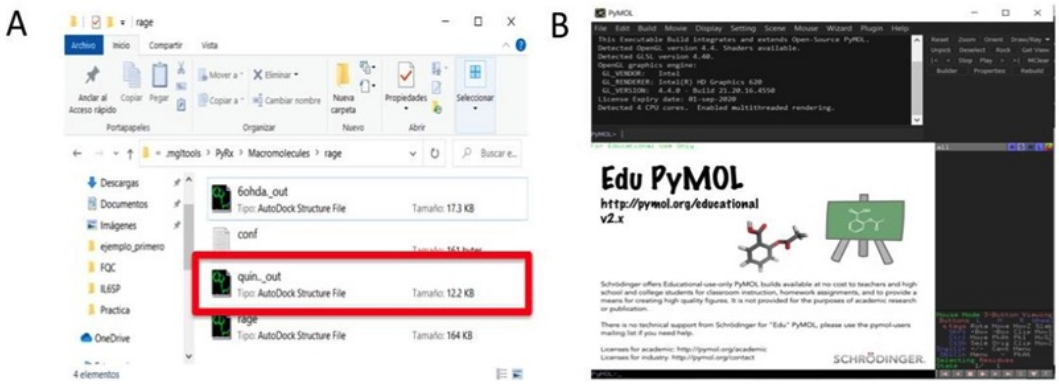
Procedimiento

1. Visualizar los confórmeros obtenidos por el estudio de acoplamiento molecular realizado en la práctica anterior a través del programa *PyMOL*.

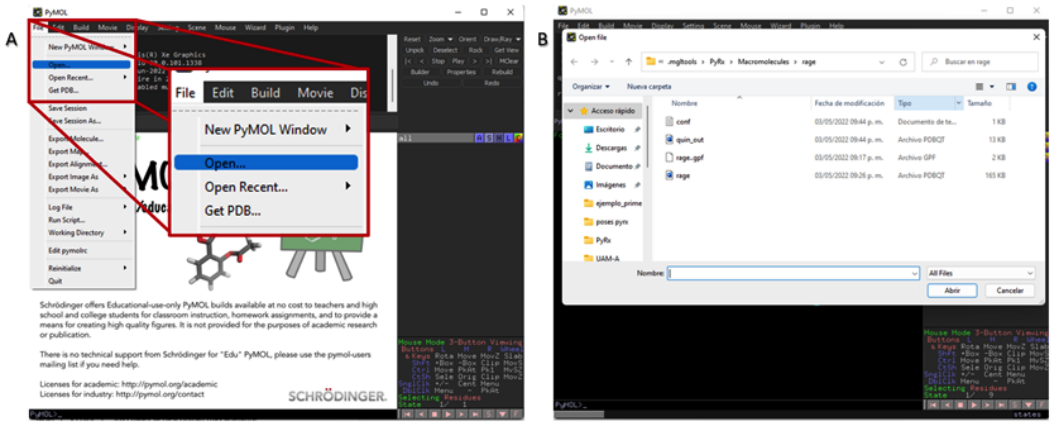
2. Exportar los confórmeros obtenidos de los estudios de acoplamiento molecular realizados en el programa *PyRx 0.8* los cuales se obtienen como se explica a continuación:

   a) Buscar la carpeta: *Mgtools -> PyRx -> Macromolecules* (normalmente se encuentran en C:-> Usuarios-> el usuario que estés utilizando por el momento). Esta última carpeta corresponde a la proteína, la cual contiene tres archivos: *protein. pdbqt*, *ligand_out.pdbqt* y *conf.txt* (en esta práctica el ligando se llama *quin_.out-pdbqt* y la proteína es "*rage*") (panel A).

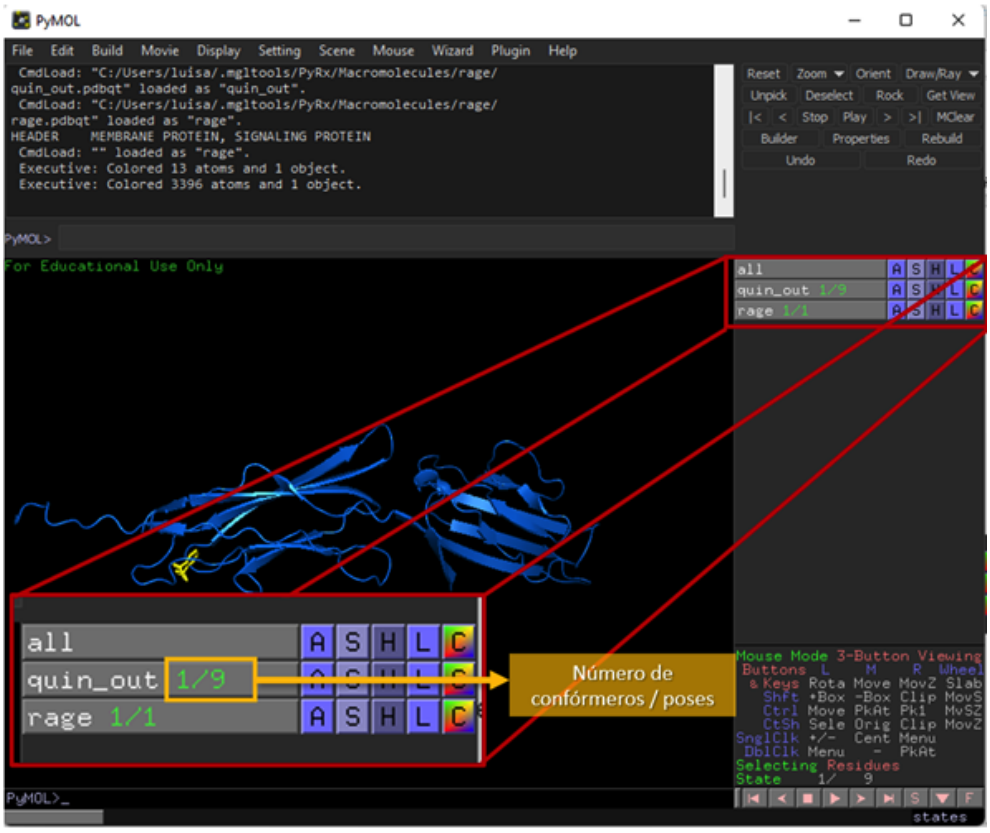
   b) Respecto al archivo llamado *ligand_out.pdbqt* contiene las mejores poses (8 o 9 orientaciones). Este archivo se puede guardar en alguna de las carpetas personales para analizar posteriormente los resultados en el programa *PyMOL* (panel B).



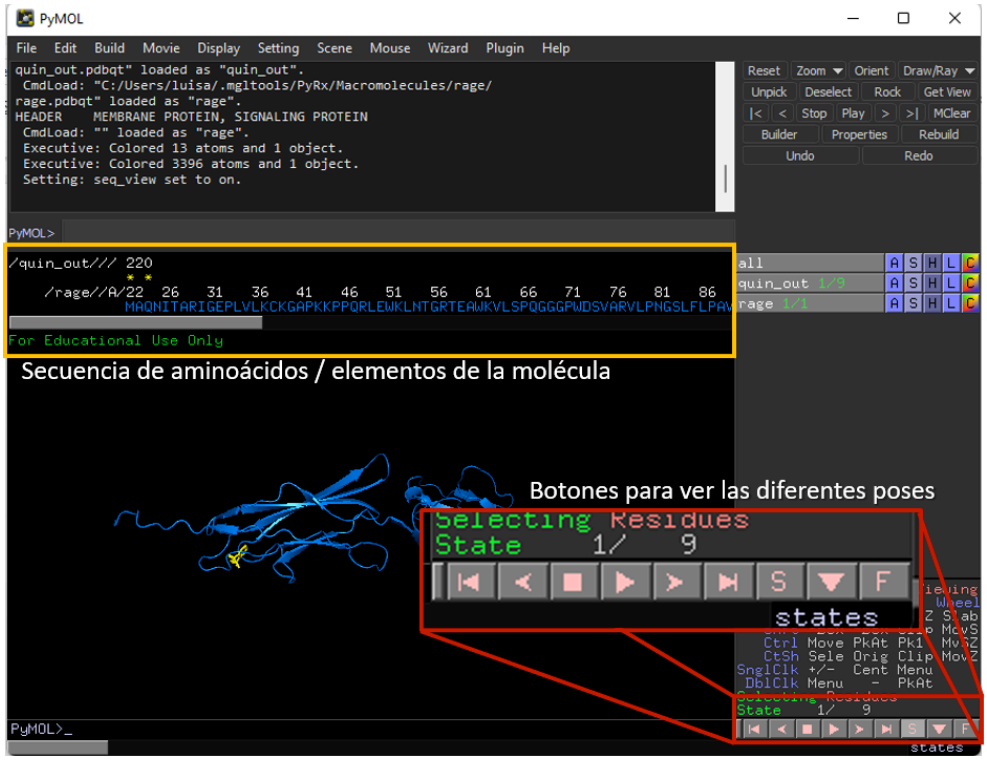
c) Al abrir el programa *PyMOL* se dará clic a "*File*" y elegirá la opción "*Open*" (panel A) y se buscará en la carpeta personal el archivo de la proteína en formato .pdb (panel B).



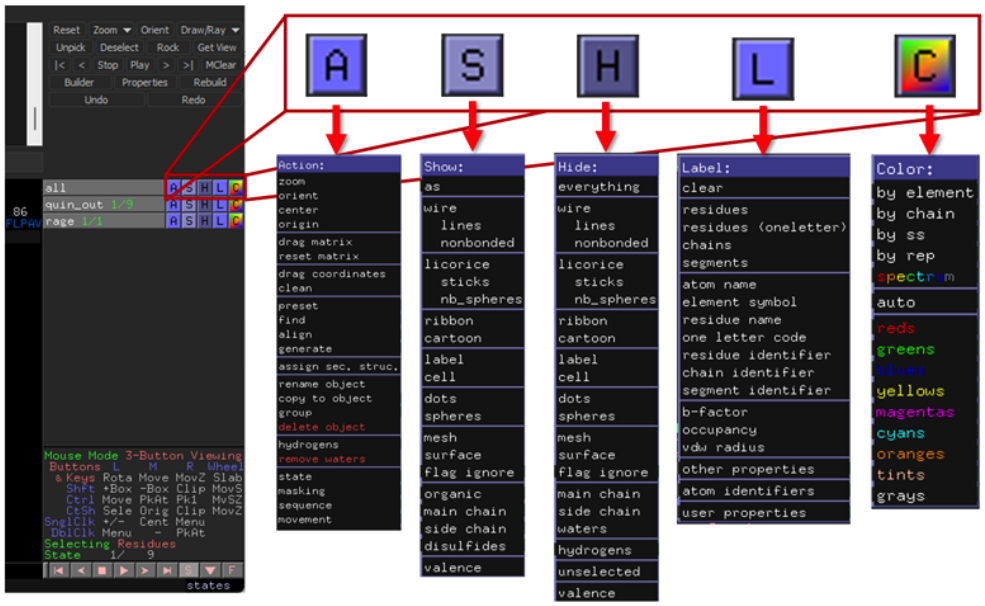
d) Después, se abre el archivo *ligand_out.pdbqt* para cargar las 8 o 9 poses. A un lado del nombre del ligando se muestra el número de la pose que se está observando. También, se registrará del lado derecho de la pantalla 5 iconos: *action* "A", *show* "S", *hide* "H", *label* "L" y *color* "C", los cuales se describirán más adelante.



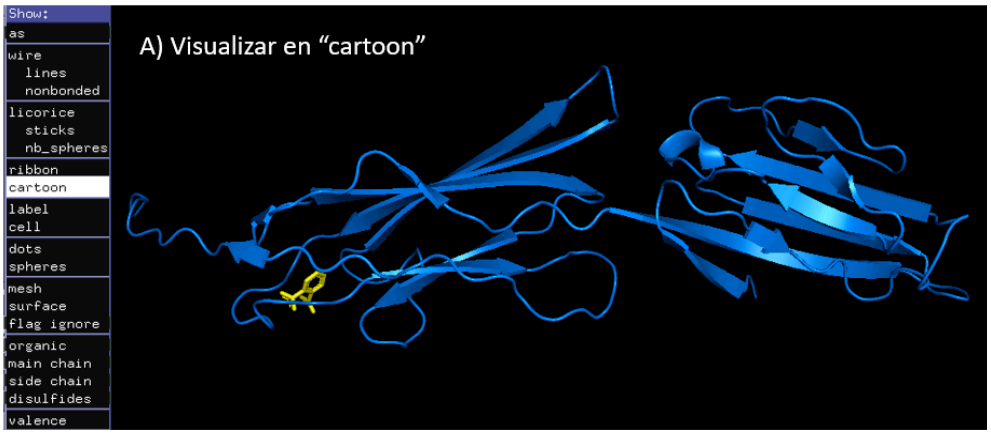
e) En la siguiente imagen se muestra en la parte inferior derecha un conjunto de botones que ayudan a visualizar las poses. El primero regresará a la primera pose sin importar cual se está visualizando, el segundo regresará a la pose anterior, los siguientes dos son para mostrar todas las poses de forma automática ("parar" y "reproducir", respectivamente), el siguiente botón llevará a la última pose sin importar cual se visualice y el triángulo hará que gire la molécula.

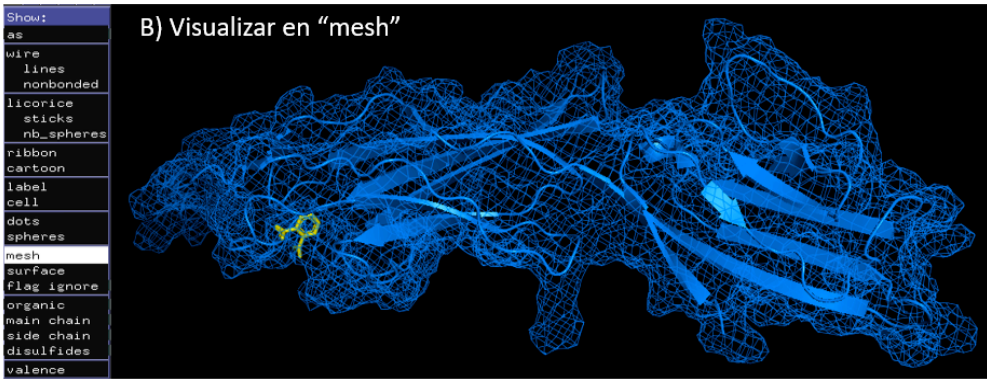


f) Como se puede observar, cada vez que se carga un archivo se registra del lado derecho de la pantalla 5 iconos: *action* "A", *show* "S", hide "H", label "L" y color "C". Las opciones de cada icono se muestran a continuación:

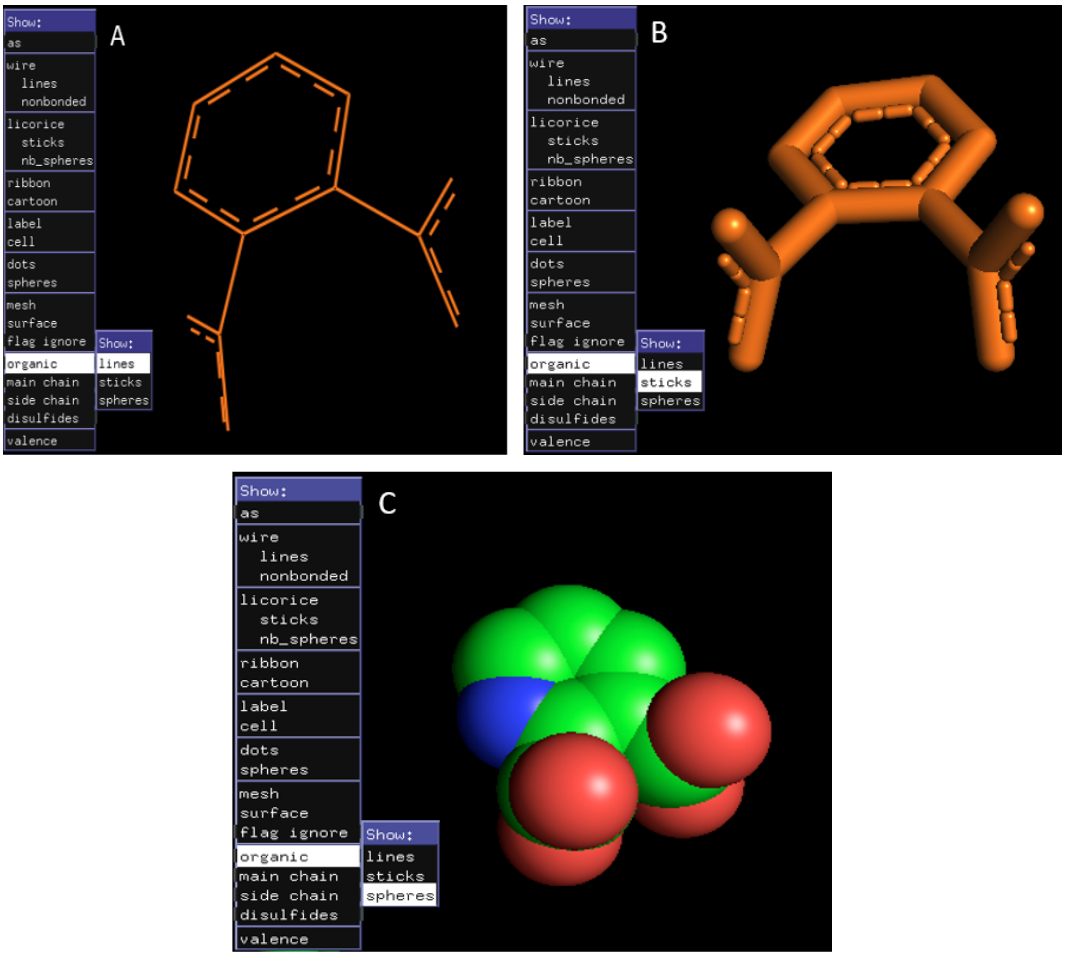


g) Estas opciones permitirán resaltar el plegamiento de la proteína dando clic en el icono "S" y después en "*cartoon*"; e inmediatamente se observará dicho plegamiento. También, se puede modificar el color de la estructura en el icono "C" seleccionando y combinando los colores, como se muestra a continuación:

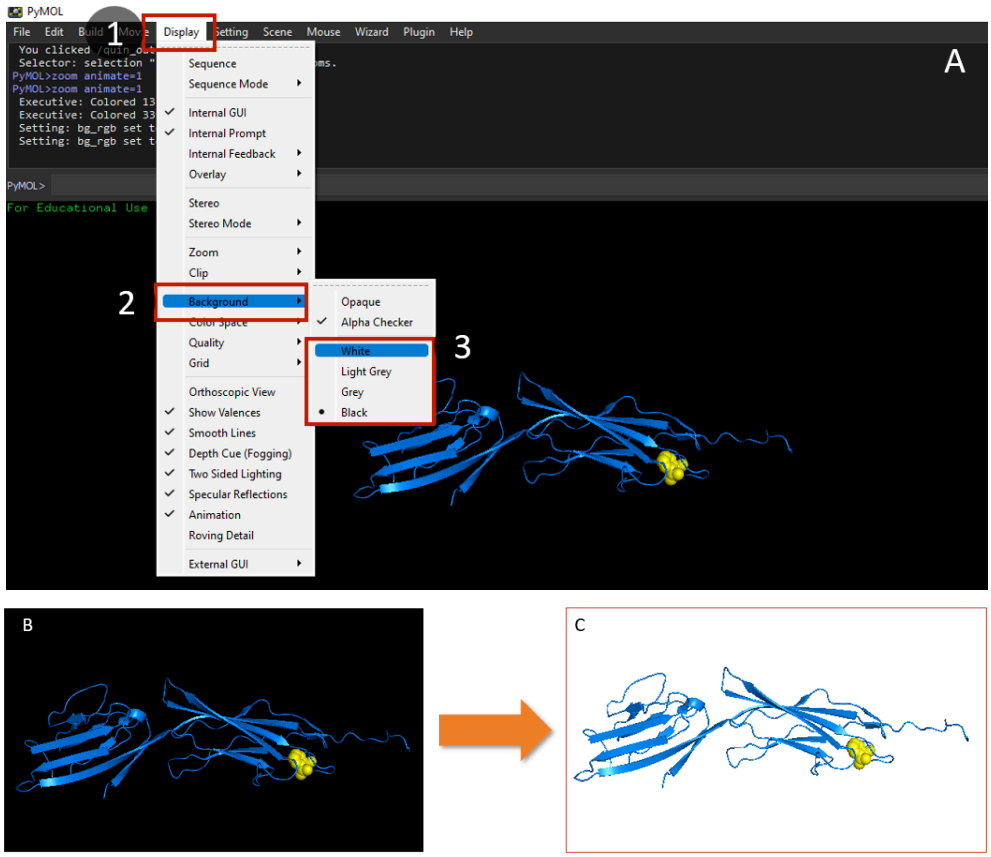




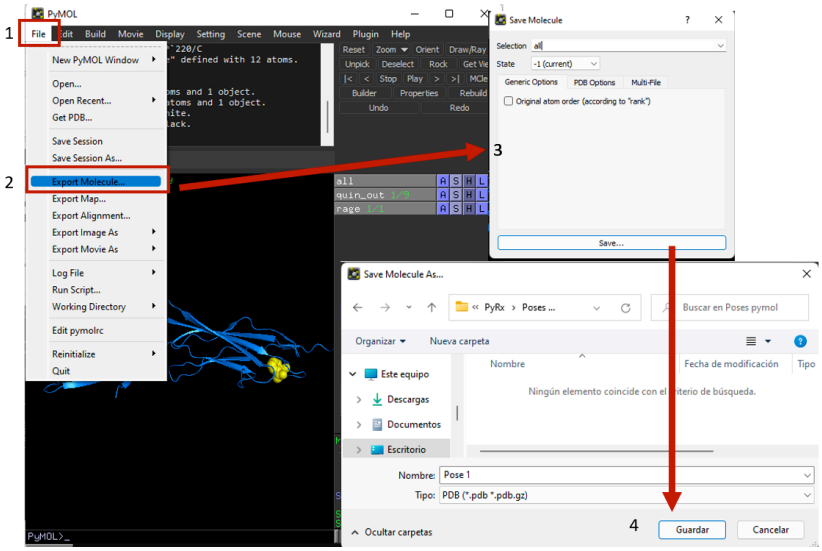
h) Para modificar visualmente el ligando, se puede seleccionar cualquier pose y se da clic en el icono "S" y después en "*organic*", se mostrarán tres opciones "*lines*", "*stick*" y "*spheres*"; es recomendable la última opción para ubicar mejor a la pose.



i) Otra recomendación importante es cambiar el fondo de la pantalla negra a blanco, para ello seleccione la opción "*Display*" y posteriormente "*Background*" -> "*White*".



j) Se generará un archivo en formato.pdb de la interacción de la proteína con la pose que se eligió en "*File*" -> "*Export Molecule*" (paso 1 y 2). Inmediatamente se despliega una ventana llamada "*Save Molecule*" mostrando la opción "*Selection*" -> "*all*"; también en la opción "*State*" -> "-1(*current*)" y "*Save*" (paso 3). Posteriormente, se elige un nombre, seleccionando el formato .pdb y "Guardar". Esto permitirá integrar un archivo que contenga a la proteína y al ligando en el número de pose que se muestra en el paso 4.



3. Finalmente se validará si el o los sitios encontrados por el acoplamiento molecular concuerdan con el sitio de unión reportado en la literatura para el sistema que se elija.

Cuestionario y/o ejercicios complementarios

1. ¿Por qué es importante usar visualizadores?

2. ¿Para qué sirve analizar un complejo o varios complejos a través de *PyMOL*?

Bibliografía

- Chapter 19. Dallakyan, S. y Olson, A. J. (2014). Small-molecule library screening by docking with PyRx. *Chemical Biology*. 243-250.
- *Part of the Methods in Molecular Biology book series* (MIMB, volume 1263).
- *The PyMOL Molecular Graphics System*, Version 1.2r3pre, Schödinger, LLC.

Casa abierta al tiempo

UNIVERSIDAD AUTÓNOMA METROPOLITANA

UNIDAD IZTAPALAPA

División de Ciencias Biológicas y de la Salud

Av. San Rafael Atlixco No.186, Col. Vicentina C.P. 09340, Del. Iztapalapa, México D.F.
Tel.: (01) 58044600