UnADM
UNIVERSIDAD ABIERTA Y A DISTANCIA DE MÉXICO

Programa de la asignatura:

Bioinformática

U3 Análisis de secuencias de aminoácidos

DCSBA

BIOTECNOLOGÍA

Índice

Presentación de la unidad

Como lo revisaste en las unidades anteriores, hoy en día contamos con una vasta cantidad de información a partir de genomas ya secuenciados de diversas especies. Esta información es útil para inferir la secuencia de aminoácidos de una proteína con base en el código genético a partir de la secuencia de nucleótidos.

En la actualidad disponemos de la secuencia de millones de proteínas basadas en la traducción teórica a través de secuencias de DNA. Es por eso que para el análisis eficiente de toda esta información se necesitan algoritmos y sistemas de cómputo que sean de utilidad para poder predecir la estructura de una proteína, su función, la interacción con ligandos y/o sus posibles ancestros evolutivos.

En esta unidad se presentan algunas herramientas bioinformáticas para el análisis de las secuencias de proteínas.

La identidad de una proteína está determinada por su estructura primaria, es decir, su secuencia de aminoácidos. Si bien los métodos experimentales que permiten la secuenciación de proteínas han sido de mucha importancia para poder caracterizar algunas proteínas, en la actualidad **la mayor parte de las secuencias de proteínas proviene de su inferencia a partir de la secuencia de nucleótidos de los genes que las codifican** (Jiménez y Merchant, 2003).

Sin embargo, es importante tener en cuenta que no siempre es posible predecir con certeza a partir de la secuencia de ácidos nucleicos las características de una proteína. En algunos casos las modificaciones que éstas sufren después del proceso de traducción son fundamentales para su función (Reece, 2004).

En muchos casos la vasta información que se tiene en las bases de datos permite hacer inferencias estructurales y funcionales con base a lo que ya se conoce para otras proteínas con secuencias similares.

Propósitos de la unidad

- Distinguir el tipo de información que se puede adquirir a partir del análisis de secuencias de aminoácidos.
- Analizar la secuencia de aminoácidos y su relación con la estructura y función biológica de una proteína.
- Identificar herramientas bioinformáticas para llevar a cabo la predicción de la estructura de una proteína y para conocer sus características fisicoquímicas.

Competencia específica

Interpretar el contenido de secuencias de aminoácidos para determinar la aplicación práctica de las herramientas moleculares a través del empleo de software.

3.1. Conversión de la información del ADN a aminoácidos

En la unidad anterior utilizaste algunas herramientas bioinformáticas para el análisis de secuencias de nucleótidos, ahora, en la presente unidad nos enfocaremos al estudio de secuencias de aminoácidos, para ello es fundamental revisar cómo podemos hacer la conversión de la información contenida en el genoma a secuencias de aminoácidos que dan origen a las proteínas.

En las bases de datos se pueden encontrar secuencias de DNA generalmente agrupadas en tres tipos: secuencias de **ADN genómico**, **ADN complementario (ADNc)** o **ADN recombinante** (Westhead, Parish y Twyman, 2002). Como su nombre lo indica el **ADN genómico** es la secuencia obtenida directamente del genoma y por lo tanto contiene genes que codifican para proteínas. Sin embargo, es importante recalcar que en el caso de los organismos eucariontes el ADN genómico incluye tanto exones, que son las secuencias codificantes, como intrones, además de elementos regulatorios y ADN intergénico. Un ejemplo es el caso del genoma humano en donde solamente el 3 % de la secuencia del ADN genómico corresponde a genes que codifican para proteínas (Reece, 2004).

El **ADNc** corresponde a la secuencia de ADN que se obtiene de la transcripción reversa a partir de ARNm, por lo tanto el ADNc constituye solamente las secuencias que codifican para proteínas (Figura 1). Una de las ventajas de utilizar la secuencia de ADNc es que debido a que proviene de la secuencia del ARN mensajero se pueden identificar genes que serían más difíciles de encontrar partiendo del ADN genómico y por lo tanto se facilita el estudio de las secuencias que codifican para proteínas.

Por último, el **ADN recombinante** corresponde a las secuencias de vectores como plásmidos que se utilizan como herramientas en la Biología Molecular y que revisaste en la unidad anterior.

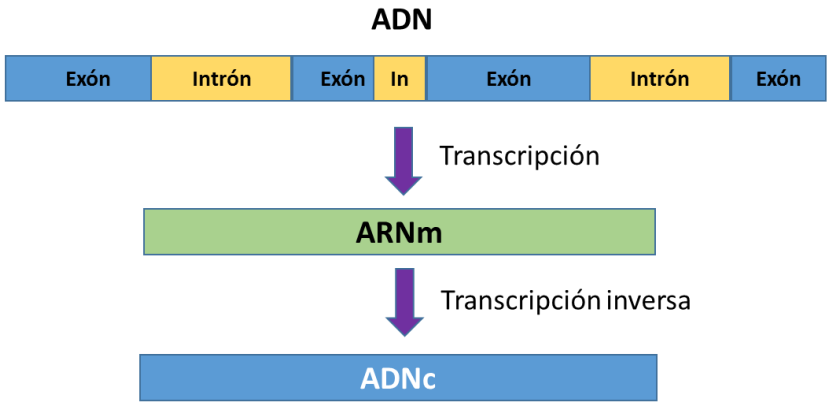


Figura 1. Representación de los procesos de transcripción y transcripción inversa.

De acuerdo al dogma central de la Biología Molecular que revisaste en la Unidad 1, la información que está contenida en el genoma, en la secuencia ADN, debe de pasar por el proceso de transcripción, para dar un ARN mensajero (ARNm) y éste después es traducido a proteínas. La manera de poder establecer una equivalencia entre la información contenida en el ADN, en el ARN y en las proteínas es el **código genético**. Mediante el **código genético** se puede explicar que las cuatro bases nitrogenadas que conforman el ADN: Adenina, Timina, Citosina y Guanina son suficientes para especificar a 20 aminoácidos con los que se construyen las proteínas. Esto es posible, ya que cada secuencia de tres nucleótidos en el ARNm constituye un codón o triplete que a su vez corresponde un aminoácido específico o a la terminación de la cadena polipeptídica. En la figura 2 se muestra el código genético, donde se observa que cada uno de los codones (tripletes de bases nitrogenadas) corresponde a un aminoácido específico. Existe un total de 64 tripletes diferentes de los cuales tres no especifican ningún aminoácido, sino que indican la terminación de la cadena polipeptídica, en tanto que el codón AUG indica el inicio además de que especifica al aminoácido Metionina.

		Segunda base					
		U	C	A	G		
Primera base	U	Phe UUU Phe UUC Leu UUA Leu UUG	Ser UCU Ser UCC Ser UCA Ser UCG	Tyr UAU Tyr UAC PARO UAA PARO UAG	Cys UGU Cys UGC PARO UGA Trp UGG	U C A G	**Tercera base**
	C	Leu CUU Leu CUC Leu CUA Leu CUG	Pro CCU Pro CCC Pro CCA Pro CCG	His CAU His CAC Gln CAA Gln CAG	Arg CGU Arg CGC Arg CGA Arg CGG	U C A G	
	A	Ile AUU Ile AUC Ile AUA Met AUG (inicio)	Thr ACU Thr ACC Thr ACA Thr ACG	Asn AAU Asn AAC Lys AAA Lys AAG	Ser AGU Ser AGC Arg AGA Arg AGG	U C A G	
	G	Val GUU Val GUC Val GUA Val GUG	Ala GCU Ala GCC Ala GCA Ala GCG	Asp GAU Asp GAC Glu GAA Glu GAG	Gly GGU Gly GGC Gly GGA Gly GGG	U C A G	

Figura 2. Código genético.

Una de las características del código genético es que es **"universal"** lo cual implica que un determinado triplete o codón lleva información para el mismo aminoácido en diferentes especies, sin embargo hay algunas excepciones en el caso de algunos protozoarios y en los códigos genéticos de organelos celulares (Hartl y Jones, 1998).

Las técnicas de ingeniería genética que permiten introducir ADN de un organismo en otro con el objetivo de que el organismo huésped sintetice proteínas de interés, son exitosas en parte porque el código genético es universal, lo que permite que, por ejemplo, se pueda llevar a cabo la síntesis de proteínas humanas en sistemas bacterianos como *E.coli.*

Otra característica del código genético es que es **degenerado**, lo cual significa que algunos aminoácidos están codificados por uno o más codones diferentes (Hartl y Jones, 1998).

Cuando se lleva a cabo la traducción del ARNm para la síntesis de una proteína, la lectura del código genético se realiza de tres en tres bases a partir de un punto de inicio. El sitio donde se inicia la lectura es muy importante ya que de esto dependerá que la traducción se realice de manera correcta. Se le llama **marco de lectura** a la sección de ADN o ARN que contiene la información para hacer una proteína completa. Una sección de lectura comienza con un codón de inicio (AUG) y termina con uno de paro (UAA, UAG o UGA), de tal manera que si se añade o se pierde un nucleótido se altera el marco de lectura y por lo tanto se modifica la secuencia de aminoácidos. Por esta razón, cuando se quiere expresar por ingeniería genética alguna proteína a partir del gen que la codifica es fundamental contar con el marco de lectura correcto.

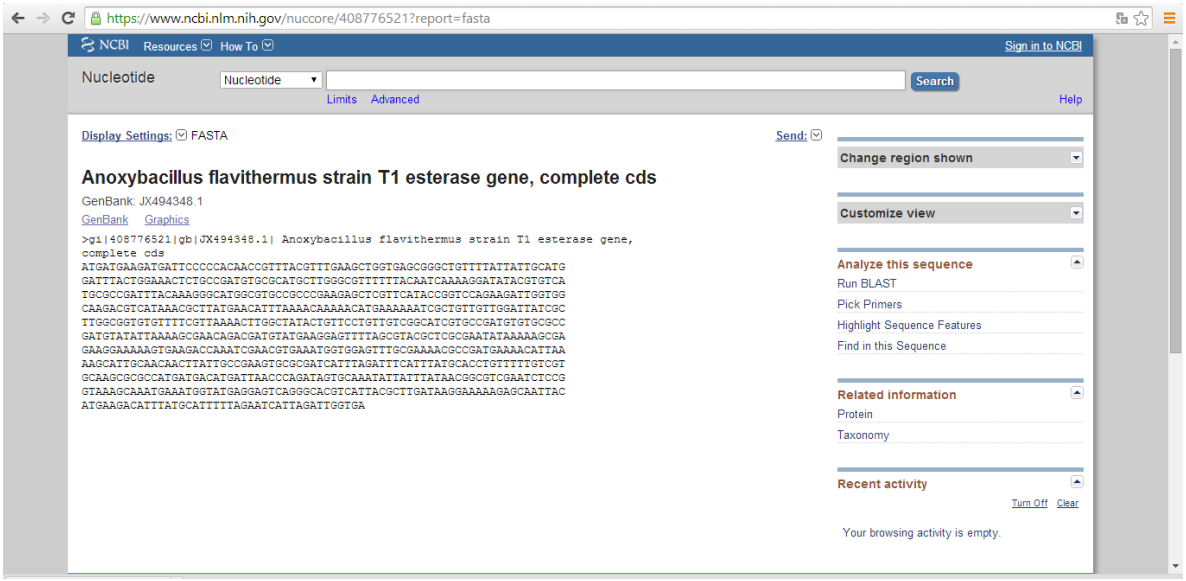
En una molécula de ADN de doble hebra, hay 6 posibles marcos de lectura, tres en dirección hacia adelante y tres en dirección reversa dependiendo de dónde se comience a leer (Brown, 2000). Si estamos leyendo una secuencia de ADN desde el primer carácter entonces el marco de lectura es +1; si se empieza desde la segunda posición entonces el marco de lectura es +2, y si se comienza desde la tercera entonces el marco de lectura es +3.

Para la secuencia complementaria, si se empieza a leer desde el primer carácter, entonces el marco de lectura es -1; -2 si se comienza desde la segunda posición y el marco de lectura es -3 si se inicia con la tercera posición. En el caso del ARN, ya que éste es de cadena sencilla solo existen tres posibles marcos de lectura. A continuación se muestra un ejemplo de como poder obtener la secuencia de aminoacidos de una proteína a partir de conocer la secuencia del ADN.

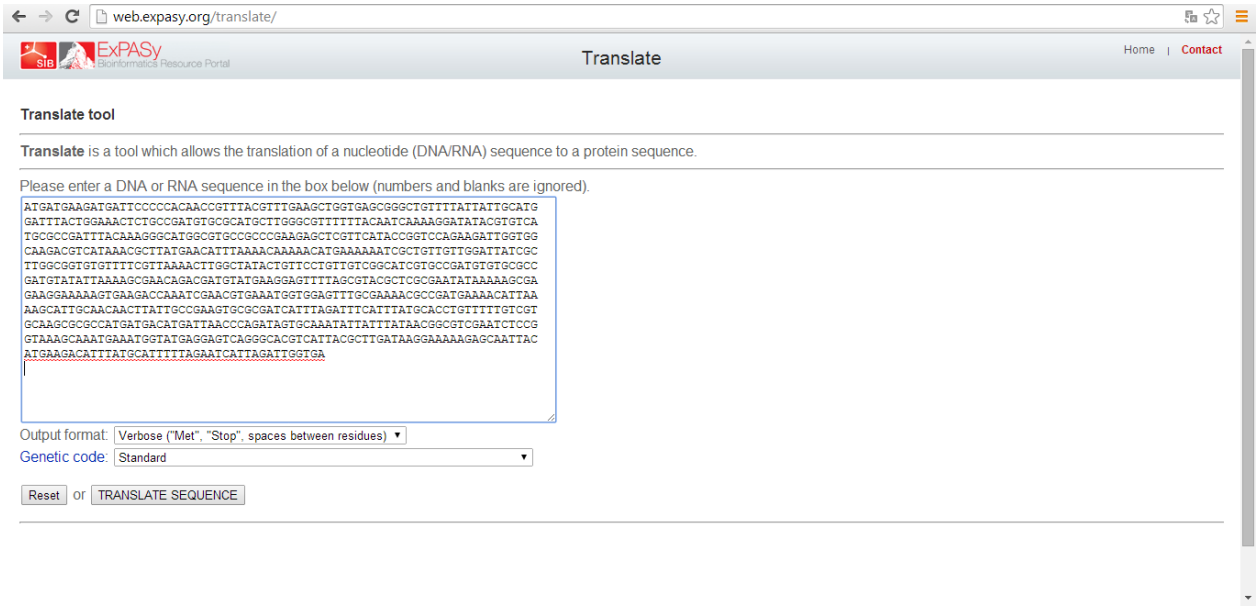
En este caso utilizaremos el ejemplo de una lipasa la cual fue aislada a partir de una cepa termofílica de la bacteria *Anoxybacillus flavithermus* (Chis *et al.,* 2013). La secuencia del gen que codifica para esta lipasa está registrada en la base de datos del GenBank con el ID: JX494348.

Ingresa a la base de datos de www.ncbi.nlm.nih.gov como lo aprendiste en la Unidad 1 y localiza la secuencia de ADN del gen de la lipasa guiendote con el ID.

Una vez que encontraste la información de la lipasa/esterasa obtén la secuencia de ADN en el formato FASTA.

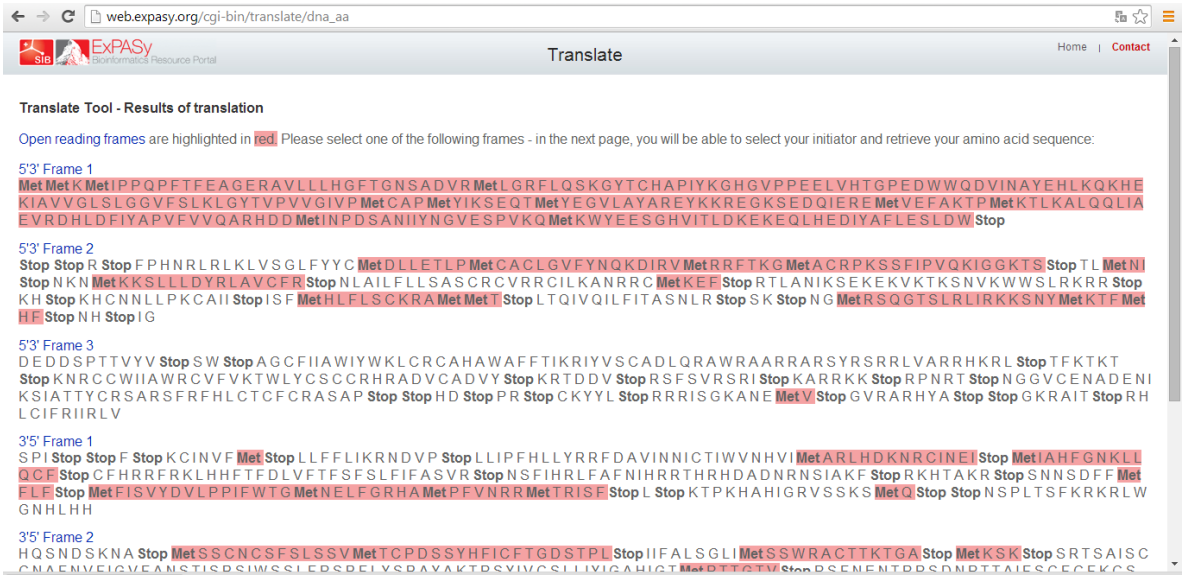


Ya que tienes tu secuencia ahora abre la herramienta de **TRADUCCIÓN** de la plataforma de EXPASY ingresando en la dirección http://web.expasy.org/translate/. Una vez abierta la siguiente interfaz copia y pega la secuencia de ADN del gen de interés.



Selecciona “TRANSLATE SEQUENCE” para hacer la traducción de la secuencia de ADN a proteína.

En la siguente interfaz aparecen las secuencias de aminoácidos y en color rosa se observan los 6 posibles marcos de lectura para la secuencia que seleccionaste. Como se mencionó anteriormente los marcos de lectura corresponden a las 6 diferentes posibilidades de lectura de la secuencia de ADN en tripletes.



Los genes que codifican para proteínas presentan un **marco de lectura abierto**, esto significa que hay un codón de inicio, normalmente ATG y un codón de paro, que puede ser TGA, TAG o TAA. Como te puedes dar cuenta en el ejemplo, en todas las opciones de lectura hay secciones en rosa que identificaron codones de inicio y codones de paro, sin embargo, el marco de lectura de la opción 1 es que corresponde al marco de lectura abierto más largo y por lo tanto el que puede considerarse que corresponde a la secuencia de aminoácidos codificada en el gen.

La lipasa/esterasa tiene 246 aminoácidos y la secuencia reportada es la que se muestra a continuación, la cual corresponde al primer marco de lectura que obtuviste y el cual es el más largo.

Secuencia de aminoácidos de la lipasa/esterasa

MMKMIPPQPFTFEAGERAVLLLHGFTGNSADVRMLGRFLQSKGYTCHAPIYKGHGVPPE
ELVHTGPEDWWQDVINAYEHLKQKHEKIAVVGLSLGGVFSLKLGYTVPVVGIVPMCAPM
YIKSEQTMYEGVLAYAREYKKREGKSEDQIEREMVEFAKTPMKTLKALQQLIAEVRDHLD
FIYAPVFVVQARHDDMINPDSANIIYNGVESPVKQMKWYEESGHVITLDKEKEQLHEDIYA
FLESLDW

3.2. Análisis de secuencias de aminoácidos

Ahora que ya sabes obtener a partir de secuencias de ADN secuencias de aminoácidos comenzaremos con el análisis, recuerda que también puedes obtener secuencias de aminoácidos directamente de diferentes bases de datos.

3.2.1. Alineamiento de secuencias de aminoácidos

Como recordarás de la unidad anterior, con los alineamientos se busca establecer una correspondencia entre dos o más secuencias. Los resultados de estos alineamientos permiten encontrar un origen evolutivo en común entre secuencias y detectar relación en la estructura y la funcionalidad de las proteínas evaluadas como lo veremos en las siguientes secciones. Además estos análisis se utilizan para generar **análisis filogenéticos**, es decir, establecer las relaciones de parentesco entre los organismos.

Los fundamentos teóricos de los alineamientos de secuencias los aprendiste en la unidad anterior, en esta sección recordaremos algunos conceptos y presentaremos particularidades para en análisis de proteínas.

A partir del alineamiento de secuencias de aminoácidos se obtiene el puntaje (score) de cada posición alineada y por lo tanto el puntaje total. En la figura 3, se ilustra la manera en que se obtiene el puntaje. Se da un puntaje de 1 si en ambas secuencias se encuentra, en la misma posición, el mismo aminoácido. Mientras que el puntaje será 0 en el caso en el que los aminoácidos sean diferentes y un puntaje de -1 se otorga cuando hay un "*espacio*" (*gap*) en el alineamiento.

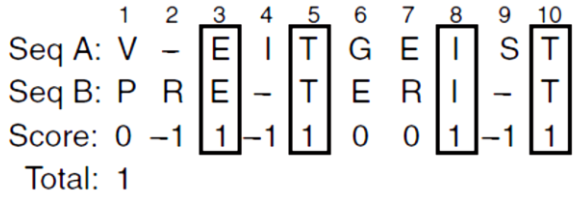


Figura 3. Obtención del puntaje a partir del alineamiento entre secuencias de aminoácidos (Adaptada de Gromiha, 2010).

A partir del alineamiento del ejemplo anterior obtenemos un puntaje de acuerdo a los aminoácidos que son idénticos en ambas secuencias analizadas. Sin embargo, en este caso no se está tomando en cuenta si las sustituciones de los aminoácidos son conservativas, es decir, si el aminoácido en una determinada posición tiene características fisicoquímicas similares al aminoácido en esa misma posición pero en la

otra secuencia evaluada. Con las **sustituciones conservativas** aunque haya cambios en los aminoácidos es más factible que se siga conservando la estructura y la función de la proteína. Es por ello que en el caso de los análisis de secuencias de aminoácidos se incorporan puntajes utilizando **matrices de sustitución**.

La matriz de sustitución toma en cuenta todas las posibilidades de sustitución de aminoácidos de acuerdo a las características de éstos. Por ejemplo, la matriz de sustitución PAM250 otorga un puntaje más alto a sustituciones que son comunes, como el caso de la sustitución de leucina por isoleucina, ya que ambos son aminoácidos hidrofóbicos. Por el contrario, los puntajes serán bajos en el caso de las sustituciones menos comunes como el caso del cambio de prolina por triptófano que son dos aminoácidos que no comparten las mismas propiedades fisicoquímicas (Twyman, 2004). Por ejemplo, el porcentaje de identidad entre la secuencia de la lactoalbúmina y la lisozima es de 38 % lo cual refleja en número de aminoácidos que son idénticos en ambas secuencias. Sin embargo, al realizar el análisis utilizando una matriz de sustitución, agrupando aminoácidos con características fisicoquímicas compartidas, el porcentaje aminoácidos similares es de 58 %.

Como veíamos en la unidad anterior BLAST es una herramienta de búsqueda para alineamientos locales, es la más usada para comparar regiones de similitud entre secuencias. En particular, el programa **blastp** compara la secuencia de aminoácidos de consulta con secuencias de proteínas en la base de datos. Con esta herramienta se pueden identificar secuencias de proteínas similares a la secuencia de consulta y localizar miembros de la misma familia de proteínas.

Con los otros programas de BLAST, **blastx, tblastn, tblastx**, se realizan las comparaciones entre la secuencia de aminoácidos y la secuencias de nucleótidos traducidos (los detalles se presentaron en la unidad 2).

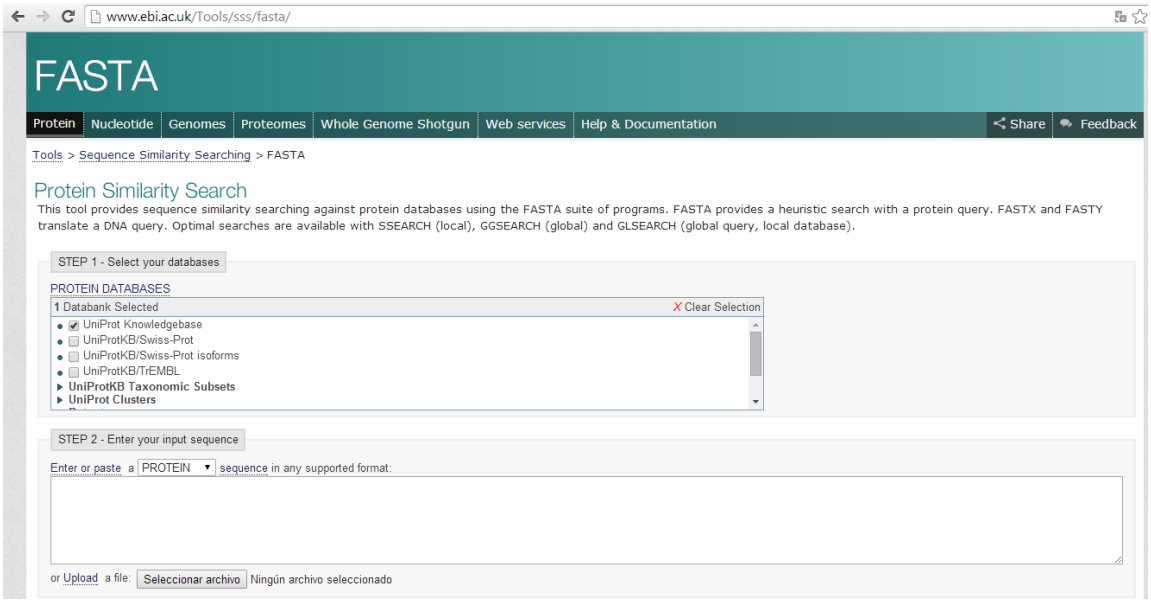
Además de BLAST otra herramienta para el análisis de secuencias es FASTA.

FASTA es un software de acceso público que sirve para analizar secuencias de ácidos nucleicos y aminoácidos. Este programa usa el método de optimización de alineamientos locales haciendo los cálculos con una matriz de substitución. La matriz de substitución describe la tasa a la cual dos secuencias divergen. Para ahorrar tiempo y costo de cómputo, FASTA hace un pre-escaneo de posibles secuencias similares y después se implementa el algoritmo de optimización. Los métodos de optimización, que buscan como comparar dos secuencias de forma que se encuentren el mayor número de similitudes, plantean una relación inversamente proporcional entre velocidad en el análisis y sensibilidad de detección, en el caso de FASTA esta relación esta mediada por el parámetro *ktup* (Schuler, 2001). Este parámetro especifica el tamaño de la unidad de búsqueda o “palabra”, aumentar el tamaño de esta unidad disminuye el ruido de fondo, es

decir hay menor cantidad de posibles similitudes, si se reduce el tamaño de la palabra, aumenta el tiempo de cómputo y hay mayor cantidad de posibles similitudes.

Tomando en cuenta que se recomienda empezar con un valor de 2 de *ktup* para buscar homología y de 1 para secuencias que se sospechan son muy divergentes (Schuler, 2001), se puede establecer la analogía entre el parámetro *ktup* y una compuerta en un rio caudaloso hecho de palabras, pequeños cambios en la apertura de la compuerta resultan en grandes cambios en la cantidad de palabras que pueden atravesarla. La siguiente fase del análisis por FASTA es usar un método heurístico que no analiza todos los posibles eventos de similitud encontrados, en cambio busca cada secuencia, con varios eventos de similitud cercanos a ella, después se les asigna valores a estas secuencias y a la secuencia con mayor puntaje se le asigna el nombre de *init1*, posteriormente se combinan varios segmentos y se forma otro valor llamado *initn*. Varias de las potenciales similitudes se contrastan con otro método que busca un nuevo alineamiento circunscrito al mejor alineamiento inicial y se le asigna el nombre *opt.* Finalmente se realiza otro análisis de alineamiento que busca todos los posibles tamaños de segmentos y optimiza la búsqueda de similitud. FASTA entrega automáticamente un valor de significancia y es importante hacer notar que los resultados entregados por el análisis dependen de la información con la cual se haga el análisis, es decir diferentes bases de datos entregarán diferentes resultados para el mejor alineamiento posible (Schuler, 2001).

Para ingresar al portal lo puedes en la siguiente dirección http://www.ebi.ac.uk/Tools/sss/fasta/. Como se muestra en la siguiente imagen en la sección de proteínas se puede hacer la búsqueda de similitud entre secuencias de aminoácidos comparando tu secuencia de consulta con secuencias de bases de datos de proteínas **Swiss-Prot, TrEMBL**.



Como recordarás en la unidad 2 vimos que **ClustalW** es un programa para llevar a cabo el alineamiento múltiple de secuencias. El alineamiento de múltiples secuencias de aminoácidos tiene muchas aplicaciones incluyendo la predicción de estructura, identificación de residuos importantes para la funcionalidad de la proteína, identificación de motivos, como lo revisaremos en las siguientes secciones.

3.2.2. Predicción de la estructura de proteína

Uno de los principales retos a vencer con el uso de las herramientas bioinformáticas es poder descifrar la estructura de una proteína a partir de su secuencia de aminoácidos. Se han desarrollado varios métodos para poder predecir la estructura secundaria y para el modelado de la estructura terciaria, a partir de la secuencia de aminoácidos.

El objetivo de conocer la estructura de las proteínas es poder obtener información de su estabilidad a partir del conocimiento de las interacciones que se pueden establecer entre los residuos de aminoácidos con el medio que rodea la proteína; así mismo, se puede inferir la localización de los residuos en la superficie o al interior de la estructura tridimensional para conocer aspectos de la funcionalidad de la proteína. Para poder entender mejor la predicción de la estructura de las proteínas repasemos los diferentes **niveles de estructuración** que se muestran en la figura 4, la **estructura primaria** corresponde a la secuencia de aminoácidos de la cadena polipeptídica. La **estructura secundaria** es un arreglo tridimensional periódico (α-hélice, β-plegada, giros), la **estructura terciaria** es la conformación tridimensional que adopta la proteína funcional "**estructura nativa**". Y la **estructura cuaternaria** involucra la asociación de más de dos cadenas polipeptídicas.

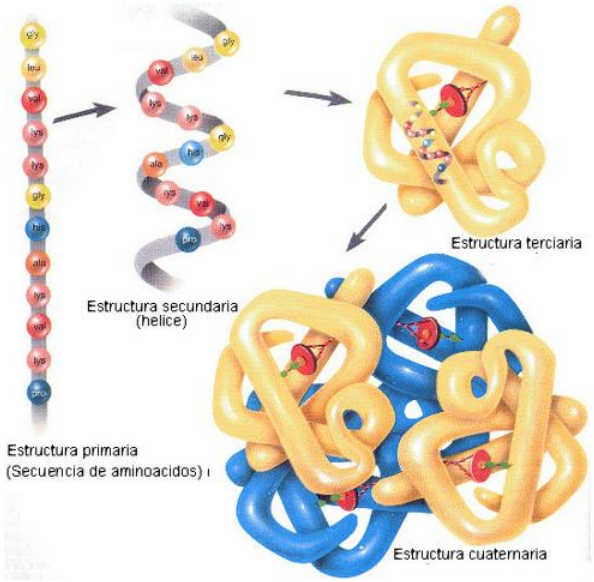


Figura 4. Niveles de estructuración de las proteínas.

Las propiedades fisicoquímicas de los aminoácidos son determinantes para poder establecer predicciones de estructura, interacciones y funcionalidad de las proteínas. Como estudiaste en la asignatura de Bioquímica las proteínas se forman a partir de 20 aminoácidos diferentes que se clasifican de acuerdo a sus propiedades en: polares, hidrofóbicos, aromáticos, cargados (negativos y positivos).

La **composición de aminoácidos** de una proteína es el número de aminoácidos de cada tipo, respecto al número total de aminoácidos (Gromiha, 2010). Se ha determinado que existe una diferencia significativa entre la composición de aminoácidos en diferentes clases de proteínas, ya que la presencia de aminoácidos específicos es fundamental para el plegamiento, estabilidad y la función de un determinado tipo de proteínas (Pautsch y Schulz, 2000). En este contexto, algunos aminoácidos tienen roles estructurales o funcionales específicos en las proteínas, como en el caso de la prolina, cisteína y glicina que se muestran en la tabla 1.

Tabla 1. Roles estructurales de algunos aminoácidos (Westhead, Parish y Twyman, 2002).

Residuo	**Función estructural**
Prolina	Al ser un aminoácido carece del grupo donador -NH para la formación de puentes de H cuando forma parte de la cadena polipeptídica. Por otra parte, a diferencia del resto de los aminoácidos, tiende a adoptar una conformación *cis* en los enlaces peptídicos. Por estas razones, la prolina interrumpe la estructura de la α-hélice, dando lugar a un punto de inflexión en la dirección de la estructura.
Glicina	La cadena lateral de la glicina está conformada solamente por un átomo de H, por lo tanto la presencia de residuos de glicina confieren una mayor flexibilidad a la cadena polipeptídica.
Cisteína	Participa en la formación de puentes disulfuro (S-S) con otra cisteína. Estos enlaces son covalentes y se forman entre residuos de cisteínas que se encuentran en diferentes partes de la cadena polipeptídica. Los enlaces disulfuro tienen contribución en la estabilidad de la estructura de la proteína.

Los métodos que se han desarrollado para la predicción de la estructura secundaria de las proteínas a partir de su secuencia de aminoácidos se basan en la **composición de aminoácidos**, sus **características fisicoquímicas** (como el perfil de hidrofobicidad), su **posición dentro de la secuencia** (Eisenhaber *et al.,* 1996; Chou y Elrod, 1999; Zhang *et al*., 2001). Además estos métodos se apoyan en el alineamiento de la secuencia de proteínas homólogas con estructura tridimensional conocida, cuando es posible.

El análisis del **perfil de hidrofobicidad** a partir de la secuencia de aminoácidos es fundamental para el entendimiento del plegamiento de la proteína y su estabilidad.

También este análisis es útil para conocer la periodicidad en la distribución de residuos de aminoácidos en regiones internas y externas en la conformación de la proteína y regiones asociadas a membranas.
En las gráficas de la figura 5 se presentan los perfiles de hidrofobicidad local de cada aminoácido en función de su posición en la secuencia de diferentes estructuras secundarias (Gromiha, 2010). Con los perfiles de hidrofobicidad se puede hacer la predicción de la estructura secundaria de acuerdo a los siguientes supuestos:

- Las **α-hélices** se localizan cercanas a la superficie de la proteína y muestran una tendencia periódica que origina regiones con alta hidrofobicidad seguida de regiones con baja hidrofobicidad (Figura 5a).

- Las **β-plegadas** (**internalizadas**) presentan agregación de aminoácidos hidrofóbicos en la región del interior de la estructura de la proteína (Figura 5b).

- Los **giros** aparecen en sitios de la cadena polipeptídica en donde la hidrofobicidad es mínima (Figura 5c).

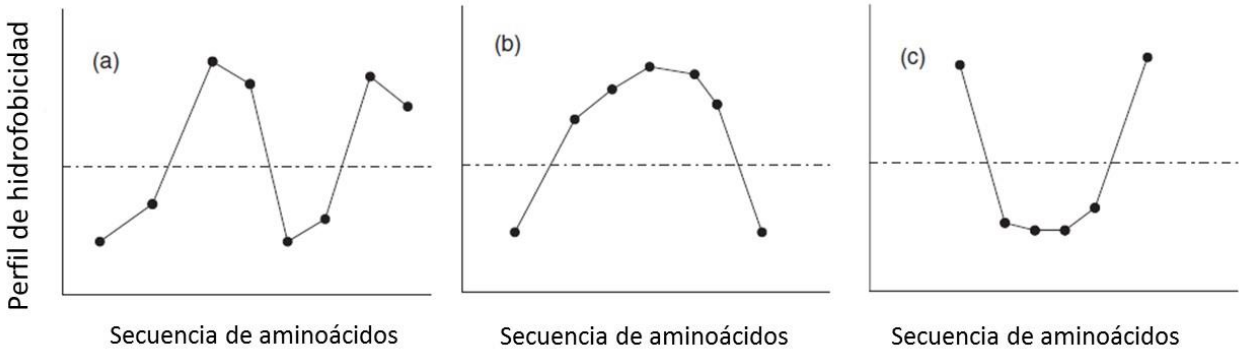
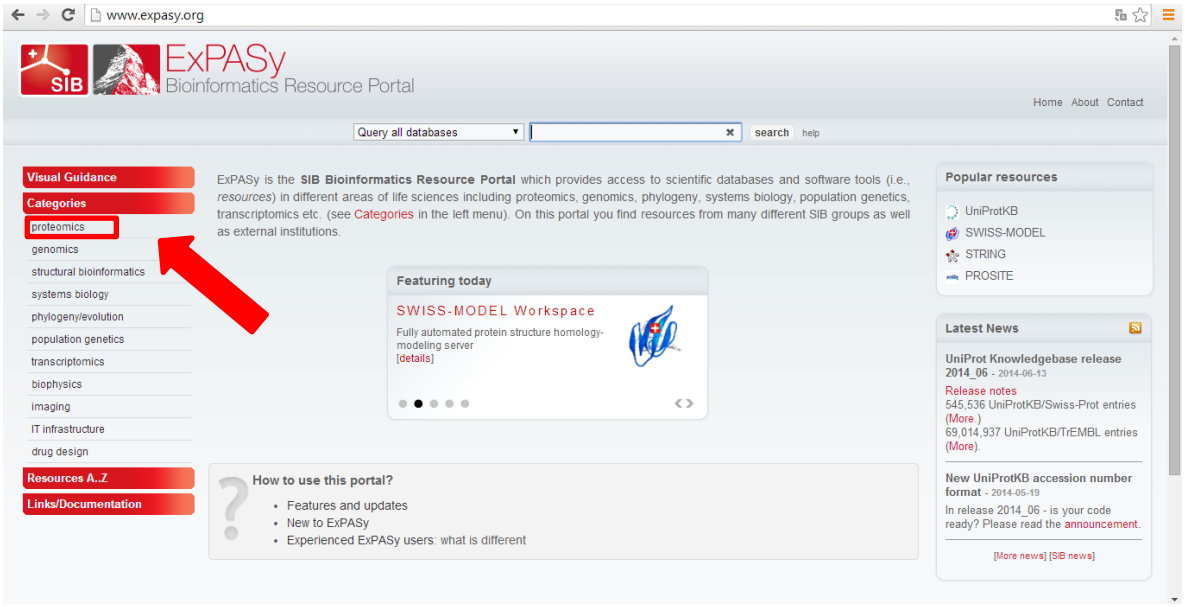


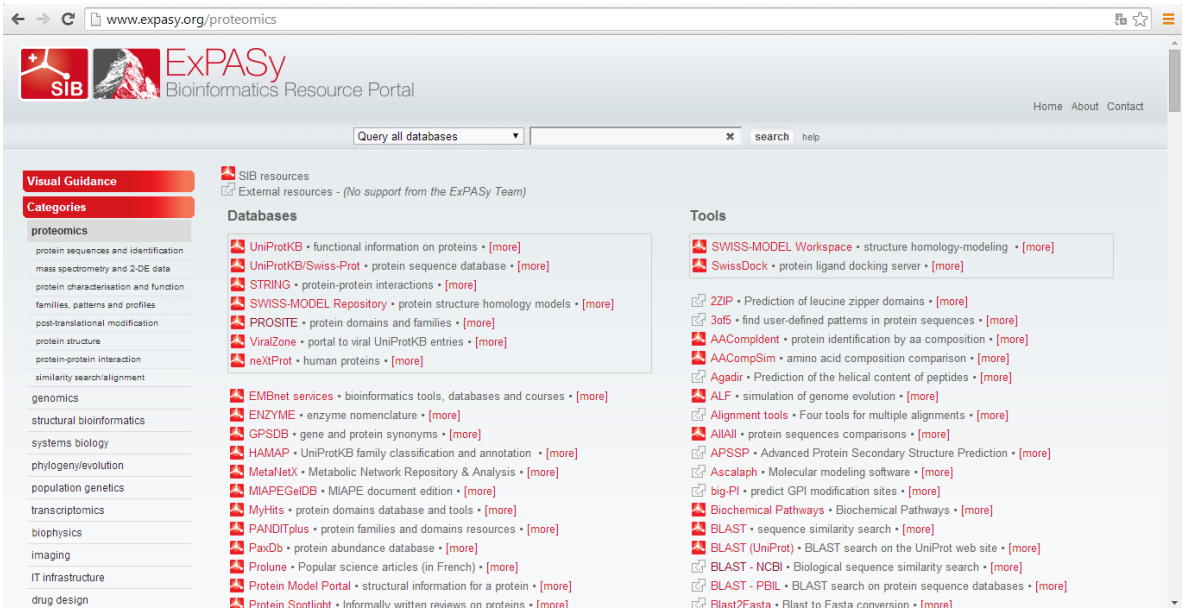
Figura 5. Perfil de hidrofobicidad de (a) α-hélice, (b) β-plegadas y (c) giro. (Adaptada de Cid et al., 1992 y Gromiha, 2010).

A través del portal Portal de Recursos Bioinformáticos **EXPASY** (Bioinformatics Resource Portal) puedes acceder a distintas bases de datos, programas y herramientas bioinformáticas para el análisis de secuencias de aminoácidos.

Entra al portal en la siguiente dirección de internet www.expasy.org.



Ahora, realizaremos la predicción de la estructura secundaria de una proteína. Entra en la categoría de proteómica (Proteomics).



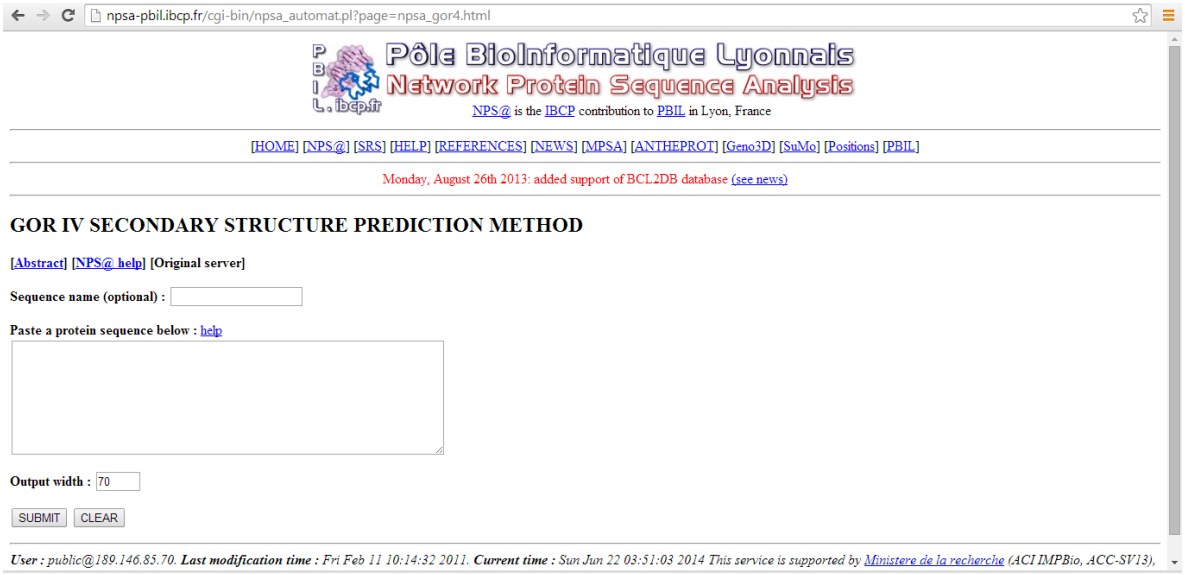
Encuentra la herramienta **GOR** con la cual haremos el análisis de estructura secundaria de la siguiente proteína GPCR. Que es un receptor asociado a proteínas G.

G protein-coupled receptor [Bos taurus]

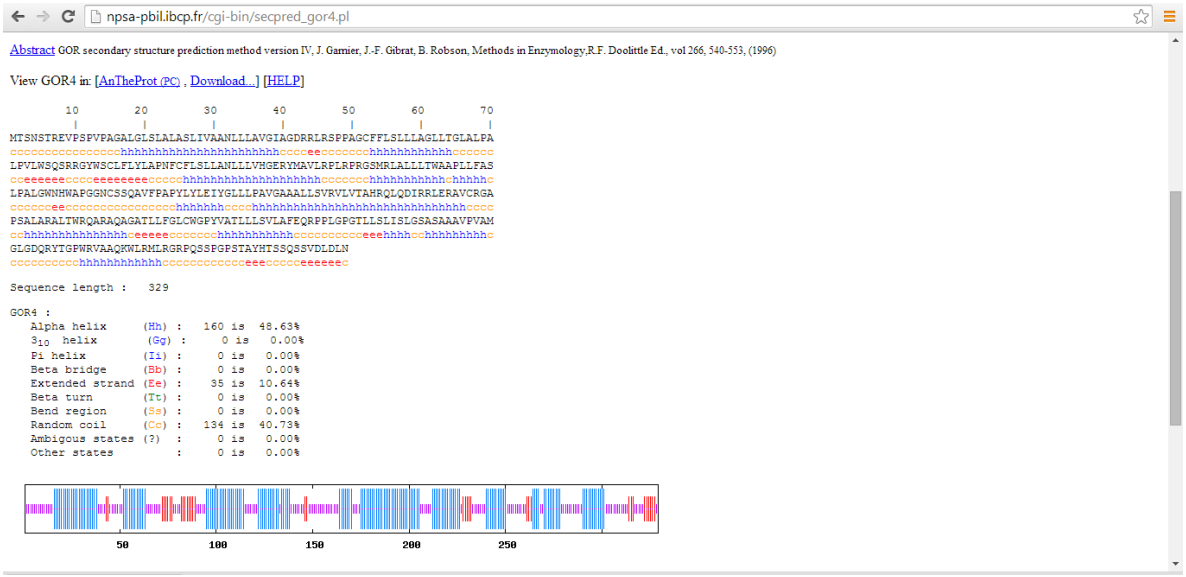
GenBank: BAC55234.1

>gi|27807551|dbj|BAC55234.1| G protein-coupled receptor [Bos taurus]
MTSNSTREVPSPVPAGALGLSLALASLIVAANLLLAVGIAGDRRLRSPPAGCFFLSLLLAGLLTGLAL
PALPVLWSQSRRGYWSCLFLYLAPNFCFLSLLANLLLVHGERYMAVLRPLRPRGSMRLALLLTWAA
PLLFASLPALGWNHWAPGGNCSSQAVFPAPYLYLEIYGLLLPAVGAAALLSVRVLVTAHRQLQDIR
RLERAVCRGAPSALARALTWRQARAQAGATLLFGLCWGPYVATLLLSVLAFEQRPPLGPGTLLSLI
SLGSASAAAVPVAMGLGDQRYTGPWRVAAQKWLRMLRGRPQSSPGPSTAYHTSSQSSVDLDLN

La siguiente imagen muestra la interfaz del programa GOR. Copia y pega la secuencia de aminoácidos de la proteína GPCR.



A continuación aparece la siguiente interfaz en donde se muestran los resultados. En la primera parte se muestra, para cada aminoácido de la secuencia, la estructura ya sea una α-hélice (h), hebra extendida (e) o hebra con enrollamiento al azar (c). Además en la gráfica de la parte inferior se presenta la distribución de la estructura secundaria en toda la secuencia de aminoácidos de la proteína.



Así como el programa GOR existen otros para el análisis de estructura secundaria que también puedes encontrar en el portal de EXPASY, algunos ejemplos son: **Agadir, APSSP, CFSSP, JPred, PSIPred, SOPMA**. Puedes realizar la predicción de estructura secundaria de la misma proteína con diferentes programas y comparar los resultados.

La proteína GPCR que estamos analizando es una **proteína transmembranal**, es decir, una proteína que se localiza en la bicapa lipídica.
En general, estas proteínas son importantes ya que se ha identificado que alrededor de un 20 a 30% de las secuencias en los genomas estudiados corresponde a proteínas membranales, las cuales son muy atractivas para su estudio como blancos para agentes farmacológicos (Gromiha, 2010). Sin embargo, debido a dificultades experimentales para la purificación de proteínas transmembranales el conocimiento de la estructura tridimensional de estas moléculas es limitado, por lo que el uso de diversos programas informáticos ha sido de gran utilidad para poder llevar a cabo predicciones estructurales y estudios de funcionalidad.

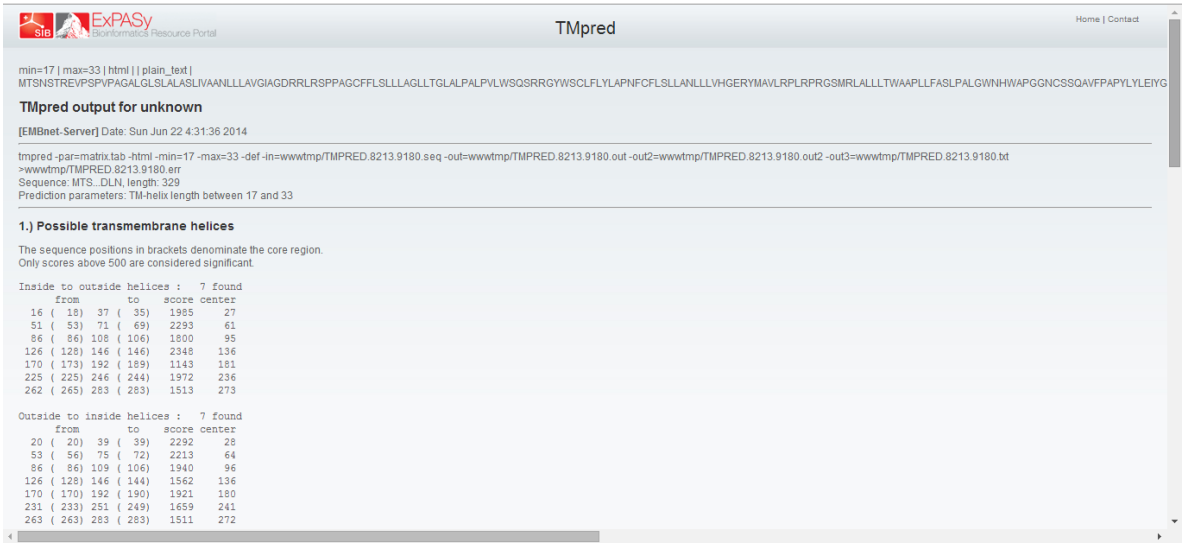
Para hacer la predicción, los programas se basan principalmente en los siguientes conceptos.

Análisis de hidrofobicidad. Debido a las características hidrofóbicas de las membranas las hélices transmembranales son predominantemente apolares y están constituidas de 12 a 35 residuos de aminoácidos (Wallin y von Heijne, 1998).

Para determinar la orientación de la proteína transmembranal se toma “**la regla del positivo adentro**”, la cual describe que la mayoría de las proteínas transmembranales tienen una distribución especial de los residuos de aminoácidos. Los residuos cargados

positivamente como arginina y lisina están ubicados en la parte citosólica (en el caso de la membrana plasmática) (von Heijne, 1986).

Ahora continuaremos realizaremos la predicción de las regiones transmembranales de la proteína GPCR con el programa **TMpred**. De nueva cuenta entra al portal EXPASY y localiza el programa TMpred, introduce la secuencia de la proteína y obtendrás los siguientes resultados.



ExPASy Bioinformatics Resource Portal

TMpred

Home | Contact

min=17 | max=33 | html | | plain_text |
MTSNSTREVPSPVPAGALGLSLALASLIVAANLLLAVGIAGDRRLRSPPAGCFFLSLLLAGLLTGLALPALPVLWSQSRRGYWSCLFLYLAPNFCFLSLLANLLLVHGERYMAVLRPLRPRGSMRLALLLTWAAPLLFASLPALGWNHWAPGGNCSSQAVFPAPYLYLEIYG

TMpred output for unknown

[EMBnet-Server] Date: Sun Jun 22 4:31:36 2014

tmpred -par=matrix.tab -html -min=17 -max=33 -def -in=wwwtmp/TMPRED.8213.9180.seq -out=wwwtmp/TMPRED.8213.9180.out -out2=wwwtmp/TMPRED.8213.9180.out2 -out3=wwwtmp/TMPRED.8213.9180.txt >wwwtmp/TMPRED.8213.9180.err
Sequence: MTS...DLN, length: 329
Prediction parameters: TM-helix length between 17 and 33

1.) Possible transmembrane helices

The sequence positions in brackets denominate the core region.
Only scores above 500 are considered significant.

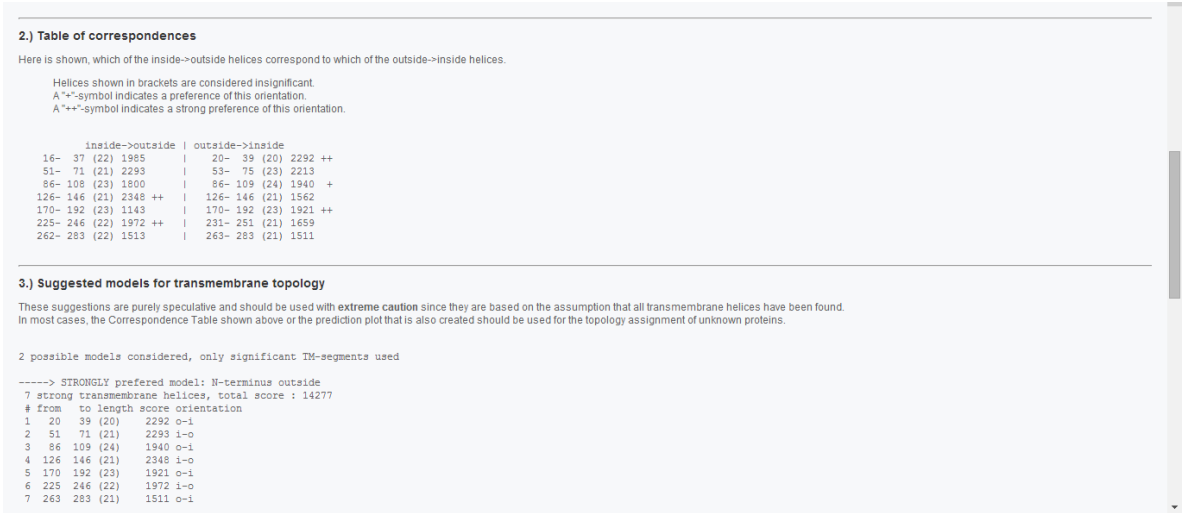
Inside to outside helices : 7 found

from		to		score	center
16	(18)	37	(35)	1985	27
51	(53)	71	(69)	2293	61
86	(86)	108	(106)	1800	95
126	(128)	146	(146)	2348	136
170	(173)	192	(189)	1143	181
225	(225)	246	(244)	1972	236
262	(265)	283	(283)	1513	273

Outside to inside helices : 7 found

from		to		score	center
20	(20)	39	(39)	2292	28
53	(56)	75	(72)	2213	64
86	(86)	109	(106)	1940	96
126	(128)	146	(144)	1562	136
170	(170)	192	(190)	1921	180
231	(233)	251	(249)	1659	241
263	(263)	283	(283)	1511	272

Los resultados de TMpred muestran que la proteína GPCR tiene 7 hélices transmembranales. Así mismo, se muestran las posiciones de los aminoácidos que están formando las 7 hélices, de acuerdo a dos posibilidades de orientación de la proteína. Si la hélice va de afuera hacia adentro de la membrana se denota como **o-i** (outside-inside).



2.) Table of correspondences

Here is shown, which of the inside->outside helices correspond to which of the outside->inside helices.

Helices shown in brackets are considered insignificant.
A "+"-symbol indicates a preference of this orientation.
A "++"-symbol indicates a strong preference of this orientation.

inside->outside	outside->inside
16- 37 (22) 1985	20- 39 (20) 2292 ++
51- 71 (21) 2293	53- 75 (23) 2213
86- 108 (23) 1800	86- 109 (24) 1940 +
126- 146 (21) 2348 ++	126- 146 (21) 1562
170- 192 (23) 1143	170- 192 (23) 1921 ++
225- 246 (22) 1972 ++	231- 251 (21) 1659
262- 283 (22) 1513	263- 283 (21) 1511

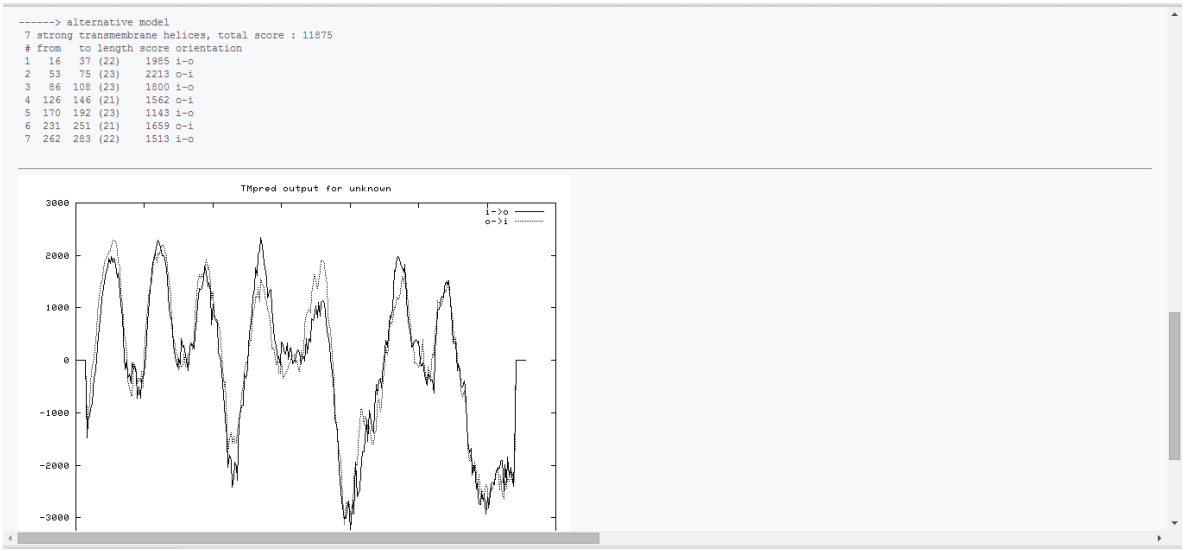
3.) Suggested models for transmembrane topology

These suggestions are purely speculative and should be used with **extreme caution** since they are based on the assumption that all transmembrane helices have been found.
In most cases, the Correspondence Table shown above or the prediction plot that is also created should be used for the topology assignment of unknown proteins.

2 possible models considered, only significant TM-segments used

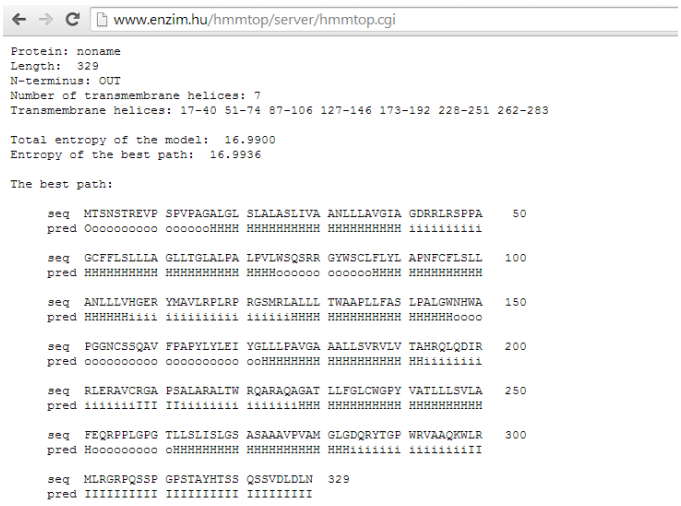
-----> STRONGLY prefered model: N-terminus outside
7 strong transmembrane helices, total score : 14277

#	from	to	length	score	orientation
1	20	39	(20)	2292	o-i
2	51	71	(21)	2293	i-o
3	86	109	(24)	1940	o-i
4	126	146	(21)	2348	i-o
5	170	192	(23)	1921	o-i
6	225	246	(22)	1972	i-o
7	263	283	(21)	1511	o-i



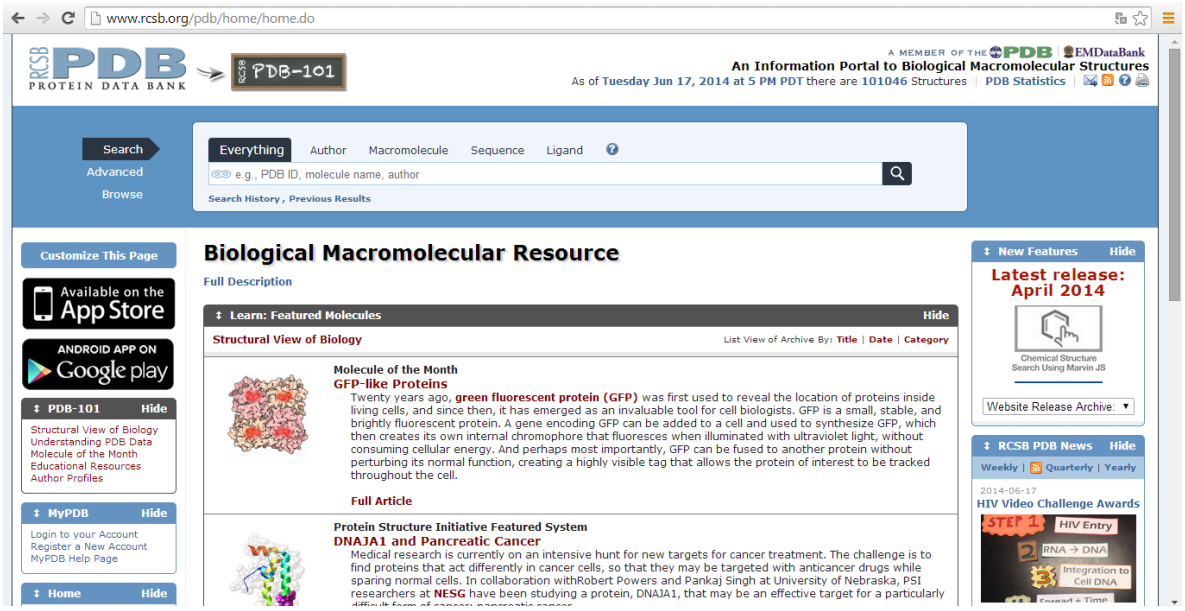
En la gráfica anterior se presenta el perfil de hidrofobicidad en donde los picos representan las regiones hidrofóbicas correspondientes a las 7 hélices transmembranales.

Finalmente en la siguiente interfaz se presenta el resumen de los resultados, en donde se observa la predicción de 7 hélices transmembranales y la posición numerada de los aminoácidos que las conforman. Además se presenta la secuencia de aminoácidos con la predicción de la topología de la proteína de acuerdo a la nomenclatura: **o**- fuera de la membrana, **H**-hélice, **i**-dentro de la membrana.



La predicción de la **conformación nativa** o **estructura terciaria** de las proteínas es un reto para la bioinformática. Para poder lograrlo se ha puesto en práctica el **modelaje**

comparativo entre la secuencia de consulta y secuencias similares de las cuales ya se conozca la estructura tridimensional por métodos experimentales. De esta manera se puede llegar a un modelo aproximado de la estructura nativa de la proteína, lo cual es muy útil, ya que como veremos en la siguiente sección el entendimiento de la estructura de la proteína es fundamental para entender su función. Con el propósito de facilitar el acceso a la información de la estructura tridimensional de proteínas y ácidos nucleicos se creó el ***P***rotein ***D***ata ***B***ank (**PDB**), el cual recopila las estructuras que se conocen hasta la fecha y que sirven de modelo para el modelaje comparativo de secuencias de proteínas de las cuales no se conoce su estructura.



La predicción de la estructura terciaria de la proteína GPCR se puede realizar en los programas **SWISS-MODEL**, **I-TASSER**. Estos programas dan un modelaje de la estructura tridimensional de la proteína de acuerdo a comparaciones entre secuencias similares de proteínas de las cuales ya se conoce su estructura.

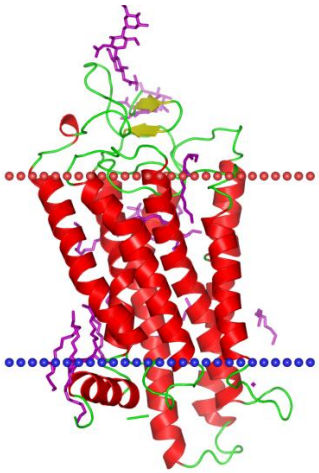


Figura 6a. Modelo de la estructura tridimensional de proteínas GPCR (tomado OPM Orientation of Proteins in Membranes, 2014 http://www.opm.phar.umich.edu).

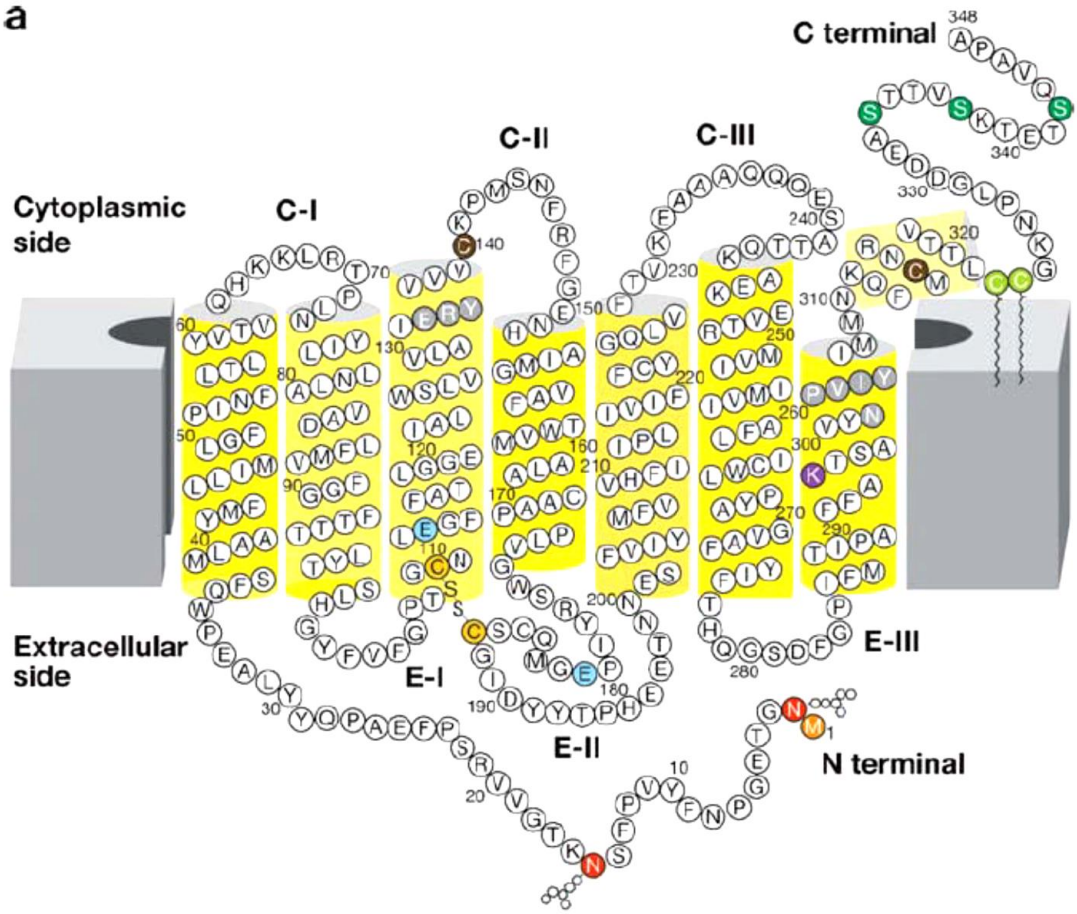


Figura 6b. Modelo de la estructura de las proteínas transmembranales GPCR (Palczewsky, 2006).

3.2.3. Relación de la secuencia de aminoácidos con la función biológica de la proteína

Como se mencionó anteriormente, la **función biológica** de las proteínas depende de su **estructura tridimensional** y de las propiedades fisicoquímicas de ciertas regiones específicas que son funcionales. Por ejemplo, el sitio activo de las enzimas es una región de la proteína donde se lleva a cabo la catálisis, por lo tanto la estructura tridimensional de esta región es fundamental para la entrada de sustratos y para que se lleven a cabo las reacciones de conversión de sustratos en productos. En el caso de otras proteínas que no son enzimas, también la conformación tridimensional es muy importante para cumplir con funciones estructurales o para el reconocimiento de ligandos y/o la interacción con otras proteínas para la formación de complejos.

Como veíamos anteriormente, cuando ocurren cambios evolutivos en las secuencias de aminoácidos estos cambios tienden a ser sustituciones entre aminoácidos con propiedades similares porque con cambios sutiles es menos probable que se altere drásticamente la estructura y la función de las proteínas. Por ejemplo, cuando se sustituye una valina es probable que el cambio sea por leucina o isoleucina que son aminoácidos que comparten características fisicoquímicas por lo que no se afectara drásticamente la estructura de la proteína (Westhead, Parish y Twyman, 2002).

Comúnmente la **función biológica** de una proteína desconocida se infiere a partir de la función de **proteínas homólogas**. Y como recordarás de la unidad anterior, la homología se determina a partir de la similitud entre secuencias, por lo tanto, en ese contexto la función de una proteína se basa en el paradigma: **secuencia similar-estructura similar-función similar** (Orengo, Jones, Thornton, 2003). Sin embargo, esto no siempre es cierto y existen casos en donde proteínas como diferente estructura, y por lo tanto diferente secuencia de aminoácidos, llevan a cabo una función similar. O bien, también existe la posibilidad de que proteínas con estructura similar catalicen reacciones diferentes, más adelante se presentaran algunos ejemplos.

Los alineamientos de secuencias son una herramienta muy poderosa para inferir la relación entre la secuencia-estructura-función de las proteínas. Normalmente en los resultados de los alineamientos se encuentran secuencias diversas con bajo porcentaje de similitud y por otro lado secuencias de **residuos conservados**, que por lo general están asociados a la preservación de estructura y función biológica que comparten las proteínas de una misma familia (Westhead, Parish, Twyman, 2002).

Una manera de resumir la información que se obtiene de los alineamientos múltiples es presentar una **secuencia consenso,** ésta es una secuencia que representa los residuos de aminoácidos comunes que se encuentran en una posición dada de un alineamiento múltiple de secuencias, un ejemplo se presenta en la figura 7. Para determinar la

secuencia consenso se toman los residuos presentes en al menos 60 % de las secuencias analizadas, si no se cumple con esta condición entonces en la secuencia consenso aparece X indicando que cualquier residuo puede ir en esa posición (Brown, 2000).

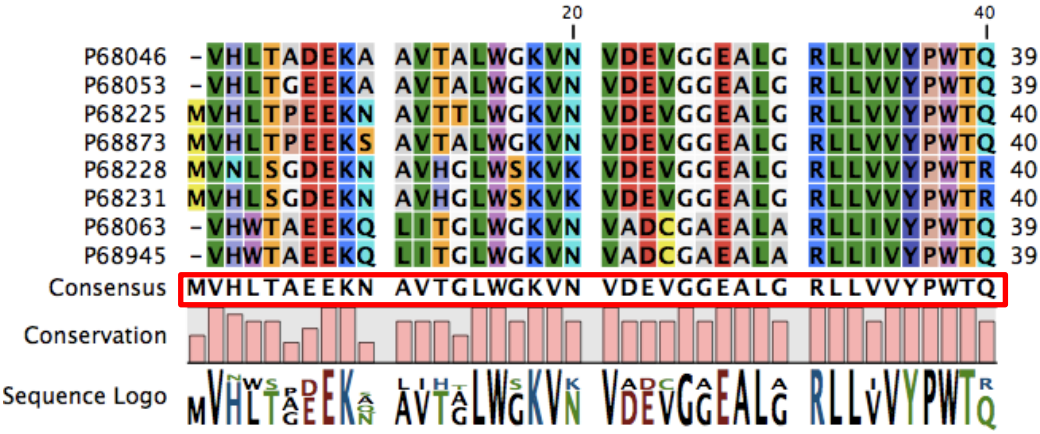


Figura 7. Se muestra la secuencia consenso (consensus) a partir del alineamiento de múltiples secuencias (tomado de Multiple Sequence Alignment, 2014 Qiagen http://www.clcbio.com/desktop-applications/features/)

Cuando un aminoácido tiene un rol específico en la funcionalidad o la estabilidad de la proteína este aminoácido es conservado en las proteínas de la misma familia. Como se mencionó anteriormente, aminoácidos como la glicina, prolina y cisteína, son importantes para la estabilidad estructural de la proteína; mientras que en el caso de las enzimas, residuos involucrados en la catálisis son esenciales para la funcionalidad de proteínas. En estos casos, los aminoácidos esenciales para la estructura y función de la proteína tienden a ser **aminoácidos conservados.** Un ejemplo se muestra en la figura 8, en donde se observa el alineamiento múltiple de las secuencias de aminoácidos de las proteasas de serina. La función biológica de estas proteínas es catalizar la ruptura del enlace peptídico de su sustrato en un sitio específico. Los residuos que son fundamentales para que se lleve a cabo la catálisis son: una serina (S), una histidina (H) y un ácido aspártico (D) los cuales se encuentran en diferentes posiciones en la secuencia de aminoácidos de la proteína pero en la estructura terciaria están muy próximos. Es lógico pensar que entre las diversas proteasas de serina estos residuos deben estar conservados ya que la reacción que llevan a cabo estas enzimas es la misma. En la figura 8 está señalado el residuo de histidina que forma parte de la triada catalítica y que está conservado en todas las secuencias analizadas. En este caso la conservación de los residuos está relacionada con la misma función entre proteínas de la misma familia, sin embargo es importante recalcar que hay ejemplos de proteínas homólogas que tienen diferente función, un ejemplo es el caso de la lisozima y la α-lactoalbúmina que se muestra a continuación (Westhead, Parish, Twyman, 2002).

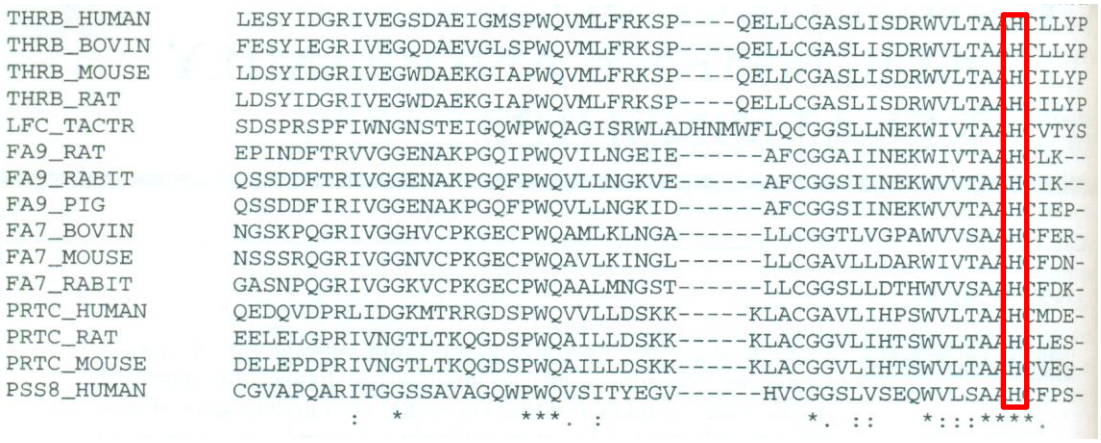


Figura 8. Sección del alineamiento múltiple de secuencias de proteasas de serina de diferentes especies. "*****" Indica un aminoácido completamente conservado, "**:**" indica un menor grado de conservación y "." Indica que solo hay algunos aminoácidos conservados (tomado de Westhead, Parish, Twyman, 2002).

La lisozima es una enzima que cataliza la hidrólisis de los enlaces β-1,4 entre los residuos de ácido N-acetilmurámico y N-acetil-D-glucosamina del peptidoglicano. Por otra parte la α-lactoalbúmina es una proteína reguladora presente en la leche producida por mamíferos. El análisis de las secuencias de aminoácidos de estas proteínas muestra que hay un alto porcentaje de identidad entre ellas, sin embargo, su funcionalidad biológica es diferente (Acharya *et al.,* 1989; Qasba y Kumar, 1997). Este es un ejemplo de evolución divergente donde los cambios por mutaciones en la secuencias de las proteínas llegan a ser significativos. En la figura 9, se muestra el alineamiento de múltiples secuencias de las proteínas lisozima y α-lactoalbúmina de diferentes especies. Como se puede observar, las regiones de la secuencia donde hay más aminoácidos conservados (*****, **:** , **.**) son los sitios con estructuras secundarias definidas y por lo general las deleciones o sustituciones de aminoácidos están en las regiones de las asas donde se afecta en menor grado la estructura de la proteína.

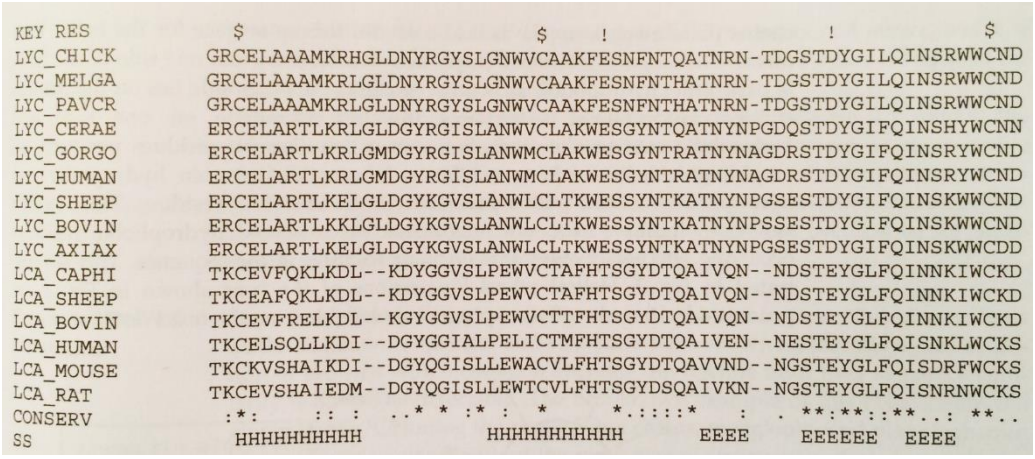


Figura 9. Alineamiento múltiple de las secuencias de aminoácidos de la α-lactoalbúmina y la lisozima (Tomado de Westhead, Parish, Twyman, 2002).

La pregunta ahora sería, **¿existe alguna relación entre el porcentaje de similitud entre secuencias de aminoácidos y la estructura tridimensional de las proteínas?**

La respuesta fue publicada por Sander y Schneider (1991), quienes encontraron que si al alinear dos secuencias de más de 80 residuos, se tiene más de 25 % de identidad, las proteínas muestran, básicamente, la misma estructura tridimensional. Existen excepciones como en el caso de proteínas que han sido manipuladas por ingeniería genética. Además de que en proteínas con menos de 80 residuos es necesario un mayor porcentaje de identidad para que se cumpla la similitud estructural.

Sin embargo, también existen casos de evolución divergente en donde aún con una muy baja similitud en la secuencia la estructura entre proteínas es muy similar. Un ejemplo es la familia de las globinas, como la hemoglobina y mioglobina. Si realizamos alineamientos entre las secuencias de estas proteínas encontramos menos del 20 % de identidad, sin embargo, estas proteínas tienen una estructura terciaria muy similar. Estas observaciones hacen evidente la importancia de conocer la estructura para inferir funcionalidad de las proteínas.

A partir de los análisis de los porcentajes de similitud entre secuencias, la estructura y la función se puede establecer si las proteínas son **ortólogas** o **parálogas**. Dos proteínas son consideradas **ortólogas** si sus secuencias de aminoácidos son muy similares, llevan a cabo la misma función en diferentes especies y han acumulado mutaciones debido a la especialización de acuerdo a la especie. Mientras que las proteínas son **parálogas** cuando de acuerdo al porcentaje de similitud de sus secuencias las proteínas son homologas, sin embargo, llevan a cabo funciones diferentes, esto se da por duplicación de genes y divergencia evolutiva (Twyman, 2004). En el caso de proteínas parálogas la predicción de la función a partir del porcentaje de similitud por alineamiento de secuencias no es confiable.

3.2.4. Análisis de secuencias específicas

La identificación de **dominios** o **motivos** conservados es de mucha utilidad para establecer relaciones entre estructura y función de las proteínas (Buehler y Rashidi, 2005). Como recordarás de la asignatura de Bioquímica, los **dominios** son unidades estructurales básicas de las proteínas. Algunos dominios tienen una función específica asociada.

Un ejemplo de secuencias específicas es la presencia del **dominio conservado** “RING finger” constituido por 40-60 residuos y que es rico en cisteínas. Este dominio está presente en células eucariontes y virales y tiene la característica de poder unir dos átomos de Zinc (Borden, 1996). Las enzimas que poseen este dominio participan en el proceso de ubiquitinación, por lo tanto es importante para la caracterización de proteínas la búsqueda de este tipo de secuencias repetidas. En la figura 10, se muestra una representación de este dominio y su interacción con los átomos de Zinc.

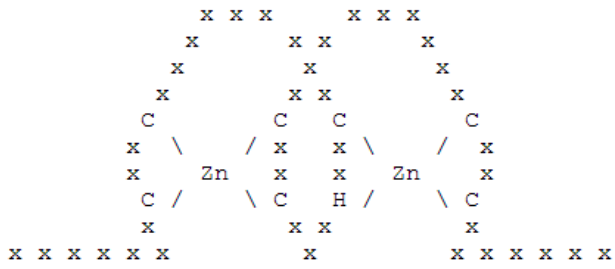
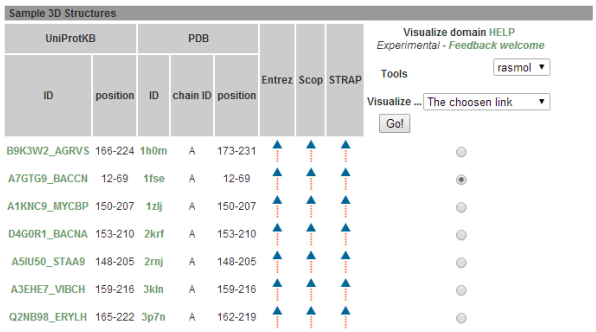
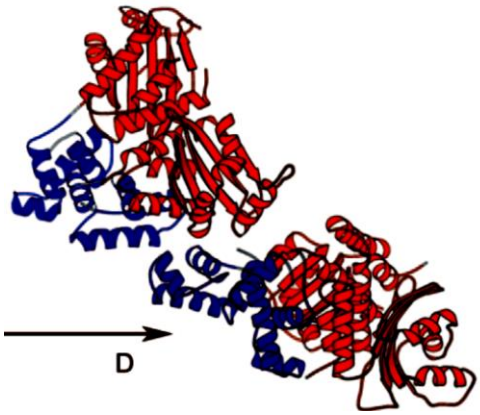


Figura 10. Representación del dominio “RING finger” en donde se muestra la interacción de las cisteínas con los átomos de zinc. (C: cisteínas conservadas asociadas a la unión del Zinc. H: histidinas conservadas asociadas a la unión del Zinc. Zn: átomo de zinc. X: Cualquier aminoácido) (Adaptado de Borden, 1996 http://prosite.expasy.org/cgi-bin/prosite/prosite-search-ac?PDOC00449

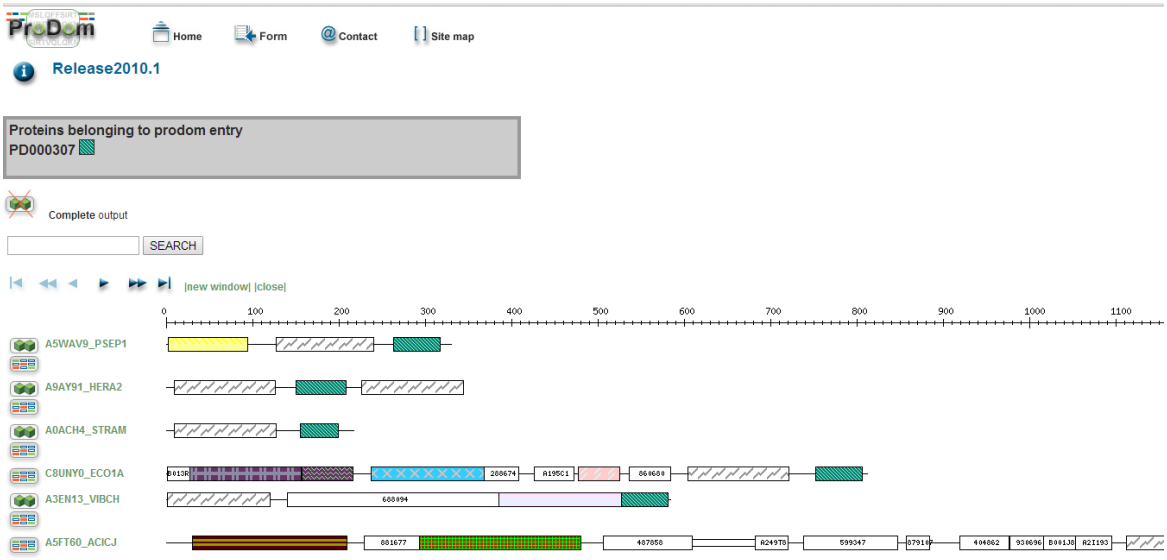
ProDom es una base de datos de dominios de familias de proteínas generada a partir de la comparación de secuencias de proteínas que ya se conocen, puedes accesar a través de la siguiente dirección http://prodom.prabi.fr/prodom/current/html/form.php.

Al ingresar tu secuencia de consulta en ProDom, el programa realiza una búsqueda comparándola con secuencias ya conocidas En la siguiente imagen se muestra el tipo de información que puedes obtener del análisis, como información general de los dominios, los nombres más frecuentes de proteínas que presentan esos dominios, características de los alineamientos y una visualización de los dominios en estructura tridimensional (Bru *et al.,* 2005).





Además de la información presentada anteriormente, ProDom da una imagen con el arreglo de la presencia de los dominos a lo largo de secuencias de varias proteínas, como se muestra acontinuación.



PROSITE es otra base de datos para el análisis de secuencias asociadas a familias de proteínas y dominios http://prosite.expasy.org/ (Sigrist, 2012).

Un **motivo** es una secuencia específica que tiene una significancia biológica. Estas secuencias repetidas pueden estar relacionadas a procesos como modificación post-traduccionales como glicosilaciones, fosforilaciones, etc. (Westhead, Parish, Twyman, 2002).

Un ejemplo es la secuencia N-{P}-[ST]-{P} que es un **motivo** para la N-glicosilación, es decir es una secuencia blanco para la modificación por glicosilación en el residuo de asparagina (N). Este motivo está conformado por una asparagina (N), seguida de cualquier aminoácido menos prolina (P), seguido de ya sea una serina (S) o una treonina (T) y al final la presencia de cualquier aminoácido menos prolina (P) (Medzihradszky, 2008). Al analizar las secuencias de aminoácidos se pueden encontrar posibles sitios de modificaciones por glicosilación, sin embargo, es este caso hay que tener cuidado con el análisis de este tipo de secuencias que son muy cortas, ya que puede a lo largo de la secuencia de aminoácidos de una proteína puede presentarse en repetidas ocasiones esta secuencia sin ser un sitio de glicosilación dando origen a un falso positivo (Westhead, Parish, Twyman, 2002).

Como te podrás dar cuenta el uso de herramientas bioinformáticas permite realizar una gran cantidad de análisis de los que se obtiene mucha información, una vez que se tienen los resultados es fundamental interpretarlos de manera correcta y en función de los objetivos que se plantearon al inicio de la investigación. Otro punto importante es que en algunos casos los resultados obtenidos con las herramientas bioinformáticas pueden

tener limitaciones y es importante verificar con ensayos experimentales cuando se quiere asegurar que una proteína cumple con determinada función biológica o tiene una estructura determinada.

3.3. Aplicaciones prácticas del análisis de secuencias de aminoácidos

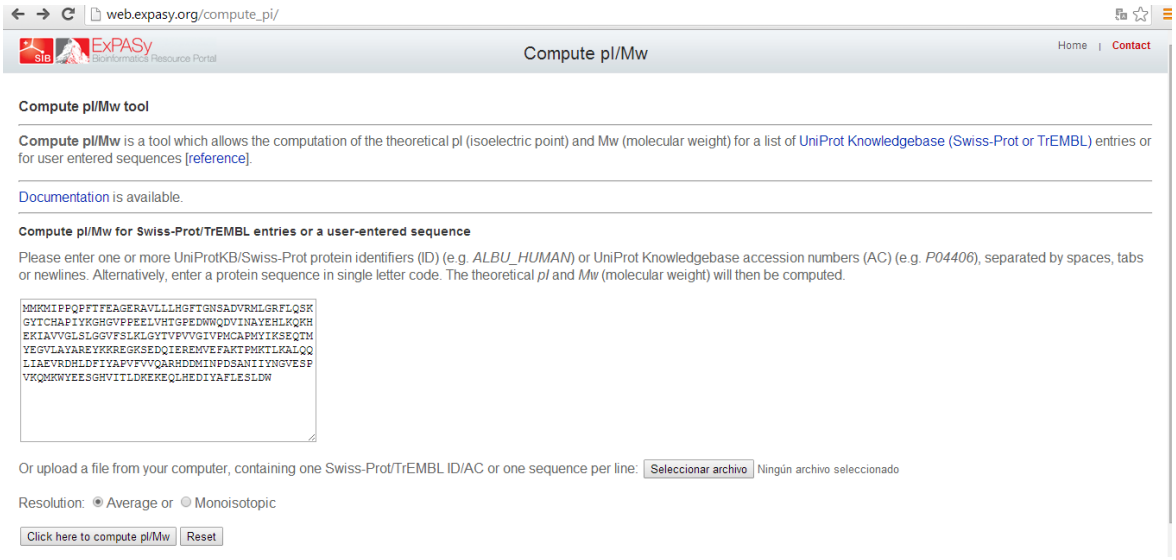
En las secciones anteriores aprendiste a utilizar herramientas bioinformáticas para el estudio de la estructura y función de las proteínas. Ahora, en esta sección abordaremos otras aplicaciones que son útiles para determinar propiedades fisicoquímicas a partir de la secuencia de aminoácidos. También tendrás un acercamiento a una fuente de información especializada en enzimas y presentaremos el **acoplamiento molecular,** el cual es un método para predecir computacionalmente la unión de una proteína con algún ligando.

3.3.1. Determinación de propiedades fisicoquímicas de las proteínas

La determinación de propiedades de las proteínas como su **masa molecular, punto isoeléctrico (pI)**, **estabilidad** a diferentes temperaturas y a diferentes valores de pH es fundamental para llevar a cabo procedimientos experimentales de purificación de proteínas y estudios de funcionalidad, por esta razón se han desarrollado herramientas mediante las cuales se pueden calcular algunos parámetros de manera teórica a partir de la secuencia de aminoácidos.

Conocer la **masa molecular** de una proteína es útil para llevar a cabo técnicas de purificación como cromatografía de exclusión molecular y para ubicar la banda correspondiente a la proteína de interés en geles de poliacrilamida.
El **punto isoeléctrico** de una proteína corresponde a un valor de pH al cual la proteína tiene una carga neta de cero ya que las cargas positivas de los residuos de aminoácidos igualan a las cargas negativas. Conocer el punto isoeléctrico y el número de residuos cargados de las proteínas es útil para llevar a cabo técnicas de purificación de proteínas como cromatografía de intercambio iónico.

Para la determinación del punto isoeléctrico y la masa molecular de las proteínas a partir de su secuencia de aminoácidos ingresa a la siguiente dirección http://web.expasy.org/compute_pi/, encontrarás una pantalla como la que se muestra a continuación. Ingresa la secuencia de aminoácidos de la lipasa/esterasa que habíamos analizado en la sección 3.1 de esta unidad.



Los resultados, que se presentan en la siguiente interfaz, muestran la secuencia de aminoácidos de la proteína y en la parte inferior se presenta el punto isoeléctrico teórico calculado que es de 5.59 y se muestra también el peso molecular teórico calculado de 28130.4 Da.

Para obtener información específicamente de enzimas puedes accesar a **BRENDA** http://www.brenda-enzymes.info/, un sitio de libre acceso donde se compila una colección de datos en relación a estructura, función y diversos parámetros de las enzimas.

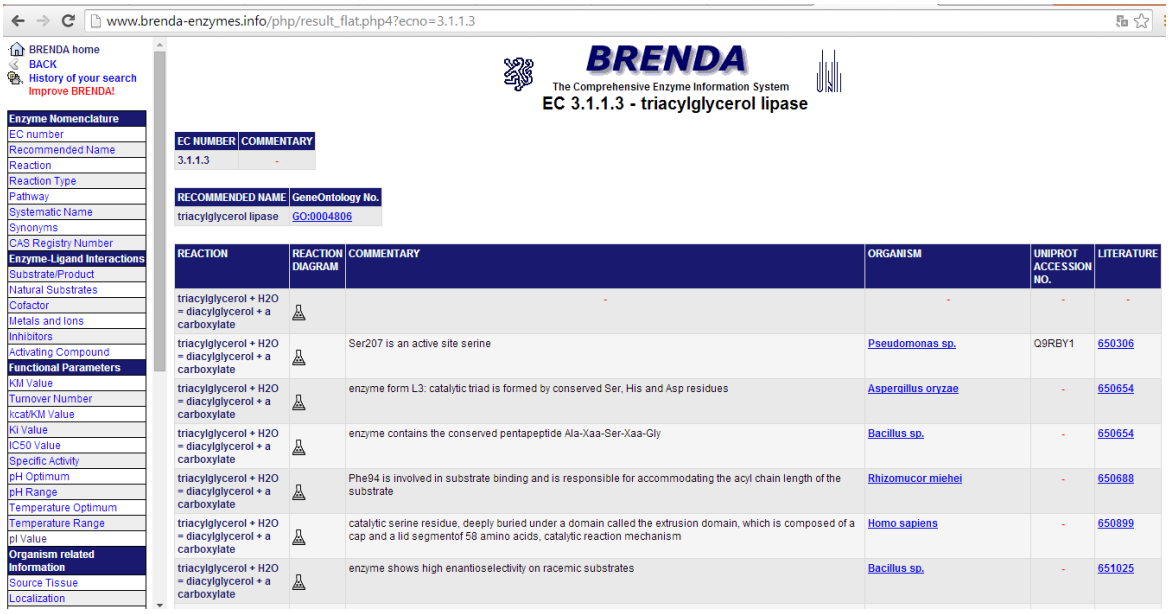
En la siguiente interfaz se presenta la información que se puede obtener en este sitio. En la sección de **nomenclatura** (*Nomenclature*) podrás encontrar una enzima de acuerdo a su nombre común, su nombre sistemático, su número de clasificación.

En la parte de **reacción y especificidad** (*Reaction & Specificity*) podrás encontrar información de la reacción catalizada por una enzima, los sustratos y productos; la ruta metabólica en la que participa, así como interacción con ligandos, inhibidores y/o cofactores.

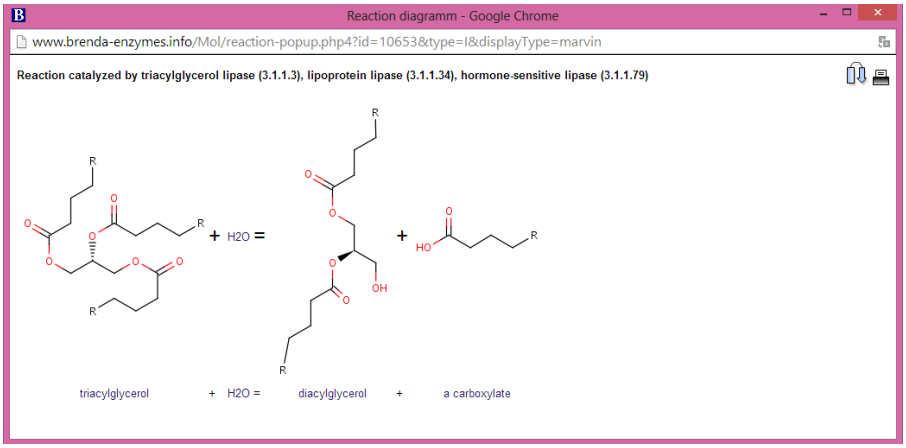
La sección de **parámetros funcionales** (*Functional parameters*) integra información de cinética enzimática como son los parámetros cinéticos: Km, número de recambio, constante catalítica, además de valores de pH óptimos, valores de temperatura óptimos y punto isoeléctrico (pI).

La sección de **estructura de enzimas** (*Enzyme structure*) presenta información relacionada con modificaciones post-traduccionales, subunidades de la proteína, links para el PDB e información sobre cristalización.
En las demás secciones podrás encontrar información de la localización subcelular de la enzima en un determinado organismo o tejido, así como datos de su estabilidad y expresión como proteínas recombinantes.

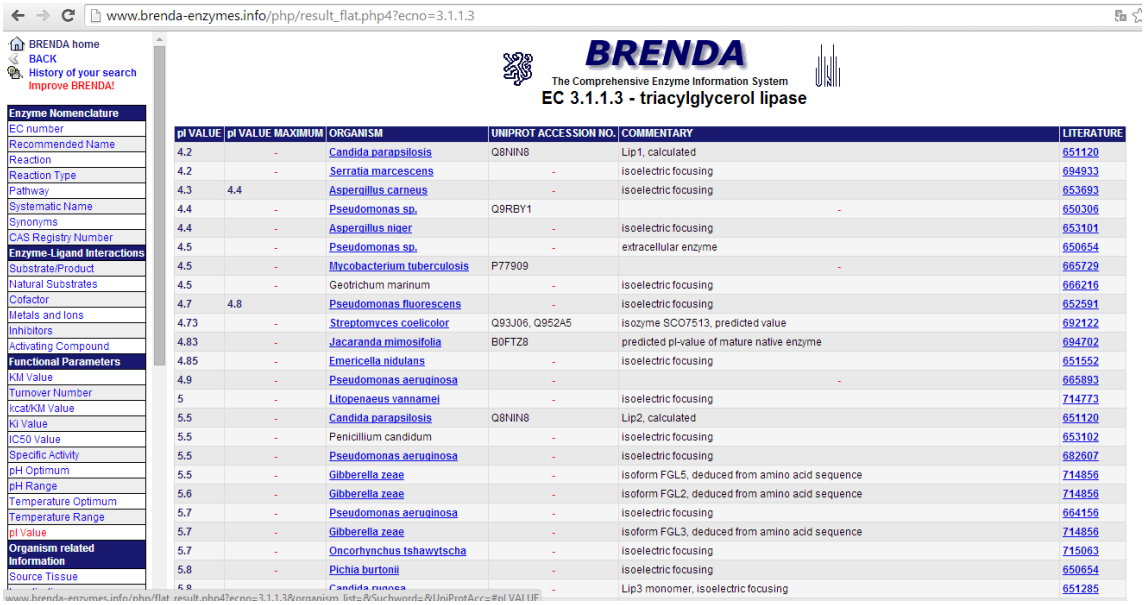
En el siguiente ejemplo se presentan los resultados para la búsqueda de información de las lipasas de triacilgliceroles, con el número de clasificación 3.1.1.3. Como puedes observar en la siguiente interfaz se presenta la reacción catalizada por la enzima, así como algún comentario si es necesario enfatizar alguna información, el organismo del cual se obtuvo la enzima y la referencia bibliográfica para obtener más información



En la sección del diagrama de la reacción (Reaction Diagram) puedes obtener un esquema, como el que se muestra a continuación, de la reacción catalizada por la enzima de interés.



También puedes accesar a valores como el punto isoeléctrico de diversas lipasas obtenidas a partir de diferentes organismos, como se muestra a continuación.



Como te pudiste dar cuenta, en este sitio se compila una gran cantidad de información de enzimas de diferentes organismos, explóralo para aprovecharlo al máximo.

3.3.2. Acoplamiento molecular para el estudio de proteínas

El **acoplamiento molecular** (**docking**) es un método para predecir computacionalmente la unión de 2 moléculas, donde una de las moléculas es una proteína. Esta aproximación es usada debido a que representa ahorro de tiempo y de dinero en las labores de investigación. Surgida principalmente de la necesidad de obtener estructuras de proteínas confiablemente y dada la dificultad de obtener información sobre las estructuras de las proteínas mediante métodos experimentales, esta aproximación computacional presenta una alternativa que tiene como principal problema perder exactitud al predecir uniones entre moléculas cuando la(s) proteínas deben pasar por cambios conformacionales para lograr unirse (Smith y Sternberg, 2002; Pérez, A., 2011).

Las interacciones moleculares juegan un papel esencial en los procesos biológicos, las interacciones involucran la interacción entre proteína-proteína, enzima-sustrato, proteína-ácido nucleico, proteína-fármaco y fármaco-ácido nucleico. Las interacciones mencionadas participan en procesos como la transducción de señales, transporte celular, regulación del ciclo celular, control de la expresión génica, inhibición enzimática, reconocimiento de anticuerpo-antígeno y la formación de complejos multiprotéicos (Hernández-Santoyo *et al.*, 2013). Para entender las uniones y la afinidad de las interacciones entre moléculas, se necesita conocer la estructura terciaria de las proteínas. Sin embargo determinar la estructura de una proteína con métodos experimentales es

frecuentemente difícil y caro, por lo cual determinar el acoplamiento de dos moléculas por métodos computacionales, se considera una aproximación importante y alternativa a los métodos experimentales, para entender las interacciones moleculares entre proteínas y otras moléculas.

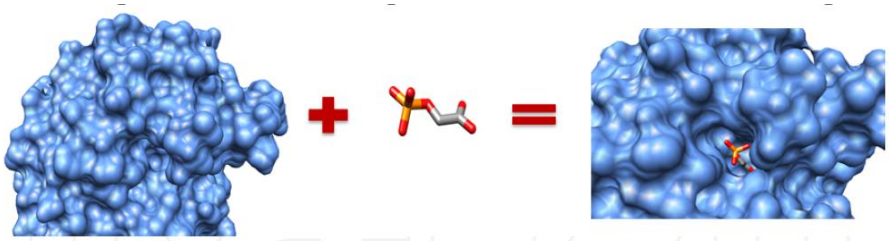


Figura 11. Elementos del acoplamiento molecular, se muestra la proteína y su ligando. (Tomada de Hernández-Santoyo et al., 2013)

La visión original de las interacciones entre proteínas es caracterizada por el modelo de llave y cerradura propuesto por Emily Fischer. Este modelo enfatiza la importancia de los efectos estéricos en la interfaz de unión para lograr especificidad de unión. En este modelo, la estructura de una proteína provee un “bolsillo” que permite a la estructura complementaria de la proteína a unirse encajar en la otra proteína. La unión frecuentemente resulta en cambios conformacionales de las moléculas que interaccionan, la especificidad es lograda mediante la adaptación de ambas moléculas mientras se unen (Pérez A, 2011).

El **acoplamiento molecular** se enfoca en predecir mediante herramientas computacionales la estructura del complejo formado por dos moléculas (Pérez, A., 2011). La importancia y eficiencia de las aproximaciones computacionales al acoplamiento de moléculas crece en medida que aumenta la descripción por métodos experimentales de la estructura de algunas proteínas. La información obtenida de la estructura de las proteínas es almacenada en bases de datos como el PDB y Worldwide Protein data Bank (wwPDK), que tienen más de 88000 estructuras de proteínas (Hernández-Santoyo *et al*., 2013). La información almacenada en estas bases de datos juega un papel crucial, ya que se puede saber la función en los procesos fisiológicos de las proteínas u obtener información sobre su afinidad de unión –usando bases de datos como AffinDB o BindDB- que como veremos más adelante, mejora la eficiencia de los procedimientos computacionales de acoplamiento.

El acoplamiento molecular como método computacional de simulación, intenta predecir la conformación de un complejo receptor con su ligando, donde el receptor frecuentemente es una proteína o un ácido nucleico y el ligando es una proteína o una molécula. La predicción adecuada de los formas de unión entre ligandos y proteínas es de vital

importancia, en un mundo que usa la estructura de las proteínas para diseñar fármacos. La aplicación más importante del acoplamiento computacional, es la selección computacional de moléculas a partir de una base de datos para investigarlas más a fondo, ya que al ser rápido y prácticamente sin costo, permite a los investigadores decidir en qué moléculas enfocar su investigación, de acuerdo a sus objetivos.

Avances significativos se han hecho en este campo, dado el incremento en la capacidad de computación, sin embargo hasta ahora poder incorporar la flexibilidad de las proteínas en la computación de las interacciones es un reto y quizá el principal problema de este campo. Se ha rebelado en estudios realizados por cristalografía de rayos X, que muchas veces hay ligandos que se encuentran unidos a una proteína y con partes de su superficie enterrada en un rango del 70 al 100%, evento que sólo puede ser logrado si la proteína es flexible (Hernández-Santoyo *et al*., 2013).

Modelar la interacción de 2 moléculas es un problema complejo, debido a la presencia muchos factores involucrados en la asociación molecular, como las interacciones hidrofóbicas, fuerzas de van der Waals, las interacciones entre amino ácidos aromáticos, puentes de hidrogeno y fuerzas electrostáticas. La modelación computacional de acoplamiento es difícil debido a que hay muchos factores que se desconocen cómo interactúan y sobretodo como afecta el solvente en el proceso de asociación entre las moléculas. Lo anterior mente mencionado termina de cobrar relevancia, cuando se toma como objetivo de este método, imitar el curso natural de interacción entre un ligando y su receptor mediante la vía de más baja energía, es decir bajo los principios de la termodinámica.

El **protocolo de acoplamiento molecular** tiene dos fases, la **función de búsqueda** y la **función de puntuación**. En la fase de búsqueda, se debe crear una cantidad óptima de configuraciones que incluyan las formas de unión ya descritas con métodos experimentales (Hernández-Santoyo *et al*., 2013). Normalmente realizar una búsqueda exhaustiva consumiría mucho tiempo y poder de computación, debido a la gran cantidad de formas en las que dos moléculas se pueden unir. En esta búsqueda se muestrea el espacio conformacional y se trata de encontrar un punto medio entre el gasto en el poder de computación y la cantidad de espacio conformacional muestreado. Algunos algoritmos usados comúnmente en la fase de búsqueda son: dinámica molecular, el método de Monte Carlo, algoritmos genéticos, algoritmos basados en fragmentos, basados en la distancia geométrica y algoritmos de búsqueda sistemática. Detalladamente la fase de búsqueda empieza con la búsqueda global de la totalidad de las configuraciones del espacio traslacional y rotacional, para identificar posibles sitios de interacción, como ya mencionamos esto consumiría mucho tiempo y poder computacional, por lo que se usan modelos simplificados de las proteínas basados en la forma geométrica de las moléculas a analizar, en las fuerzas electrostáticas y complementariedad hidrofóbica. Los modelos simplificados de proteínas “suavizan” los detalles a nivel atómico, para que puedan ser

representados como figuras geométricas discretas (figuras geométricas individuales y separables del conjunto) (Pérez, A., 2011). Sin embargo la forma geométrica no es la única restricción que determina los eventos de interacción, las propiedades químicas entre los residuos que interactúan y la energía libre son vitales para determinar la unión entre receptor y ligando. Algunos algoritmos permiten superposición estérica, para dar cuenta del cambio conformacional de la proteína, sin embargo se debe tener cuidado con este parámetro ya que puede resultar en la identificación de falsos positivos (Pérez, A., 2011).

Después de muestrear el espacio conformacional, se elige una región de interés, dado que la función de búsqueda típicamente arroja muchos candidatos, los cuales como ya vimos provienen de modelos simplificados y muchos de ellos no corresponden a la configuración nativa a nivel atómico. Una forma de solucionar este problema, es modelar los cambios de las cadenas secundarias y la estructura principal que ocurren cuando se induce la unión. En esta etapa, la gran cantidad de grados de libertad (número de elementos que son libres de cambiar, al final del cálculo) el cálculo consume muchos recursos computacionales. Por lo cual la cantidad de las estructuras candidatas se debe reducir a un número manejable (determinado por los investigadores). Una aproximación común es quedarse con las estructuras que sean más estables (porque tienen niveles bajos de energía libre de Gibbs) o un conjunto de estas estructuras. Después se agrega al proceso de selección, información sobre las propiedades físico-químicas de las regiones de unión, evidencia empírica de los residuos de contacto en la región de unión y residuos conservados, con la finalidad de incrementar la especificidad de la función de puntaje para esta etapa.

Posteriormente se intenta refinar las estructuras candidato, en esta etapa se concentra en los cambios hechos en la región de unión. Se intenta minimizar la energía libre quitando la superposición de elementos estéricos y la optimización de las interacciones estéricas y electrostáticas. Un pequeño ejemplo de cómo es el procedimiento en esta etapa usando el algoritmo de Monte Carlo, sería tomar un residuo y su isómero conformacional se cambia por otro. Se calcula la energía libre total y se mantiene el cambio si la energía libre es menor con el cambio. Este proceso se repite hasta que se alcanza la minimización de la energía libre.

Finalmente se elige un modelo, esta etapa se enfatiza la elección de estructuras que se parezcan a la estructura nativa de la proteína, por lo cual ahora no se permiten errores de correspondencia entre la estructura candidata y la estructura nativa, es decir no se permite “suavización” de las estructuras.

El otro gran componente del protocolo de acoplamiento molecular es la función de puntaje, que consiste de una serie de métodos matemáticos usados para predecir la fuerza de las interacciones no covalentes, a lo que se llama afinidad de unión.

Hay tres importantes aplicaciones de las funciones de puntaje en el protocolo de acoplamiento molecular. El primero es determinar la forma de unión y el sitio del ligando en la proteína. La segunda aplicación es predecir la afinidad de unión absoluta entre una proteína y un ligando, que es particularmente importante para mejorar la eficiencia de futuros fármacos. La tercera aplicación para el diseño de fármacos basado en la estructura de las moléculas, es identificar el potencial radio entre moléculas con posibles efectos terapéuticos para una determinada proteína. Algunas de las funciones de puntaje usadas son: campo de fuerza, empírica, basada en información previa, puntaje por consenso y complementariedad de formas. La última necesita de la descripción de la superficie de la proteína, lo cual puede hacerse transformando la estructura de la molécula en segmentos de una red (suavizarla), formando cuadrantes o se puede usar un algoritmo que calcule la superficie accesible al solvente o la superficie inaccesible al solvente.

Como mencionamos anteriormente, uno de los problemas actuales en el campo del acoplamiento molecular, es lidiar con la flexibilidad de las proteínas para unirse a su ligando. Es por esto que el acoplamiento molecular se puede dividir en rígido y flexible.

Acoplamiento rígido: esta aproximación toma al ligando y a la proteína como estructuras rígidas y las explora a nivel traslacional y rotacional con sólo 6 grados de libertad.

Acoplamiento flexible: Una aproximación común bajo este marco es modelar la unión del ligando con la proteína asumiendo que el ligando es flexible y la proteína rígida, por lo cual sólo se considera el espacio conformacional del ligando. Idealmente también se debería tomar el espacio conformacional de la proteína y hay métodos que lo contemplan (Hernández-Santoyo *et al*., 2013). Hay tres categorías generales de algoritmos para modelar asumiendo flexibilidad en el ligando: métodos sistemáticos, métodos estocásticos o al azar y métodos de simulación. Debido al gran tamaño de las proteínas y su gran cantidad de grados de libertad, modelar su flexibilidad es uno de los principales retos. Para modelar la flexibilidad de las proteínas se han desarrollado métodos que pueden ser clasificados en: acoplamiento suave, flexibilidad de cadenas secundarias, relajación molecular y acoplamiento de ensamble de proteínas.

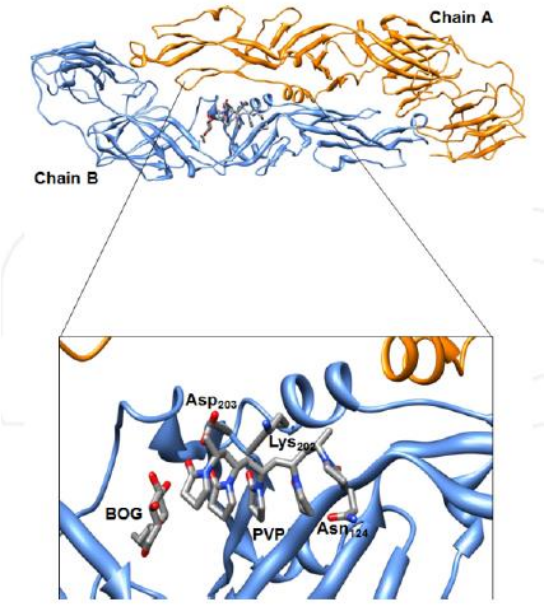


Figura 12. Acoplamiento molecular entre la proteína del virus del dengue y su ligando PVP, se muestran los aminoácidos que participan en la interacción. (Tomada de Hernández-Santoyo et al., 2013).

Actividades

La elaboración de las actividades estará guiada por tu docente en línea, mismo que te indicará, a través de la Planificación de actividades, la dinámica que tú y tus compañeros (as) llevarán a cabo, así como los envíos que tendrán que realizar.

Para el envío de tus trabajos usarás la siguiente nomenclatura: BIIN_U3_A1_XXYZ, donde BIIN corresponde a las siglas de la asignatura, U3 es la etapa de conocimiento, A1 es el número de actividad, el cual debes sustituir considerando la actividad que se realices, XX son las primeras letras de tu nombre, Y la primera letra de tu apellido paterno y Z la primera letra de tu apellido materno.

Autorreflexiones

Para la parte de **autorreflexiones** debes responder las *Preguntas de Autorreflexión* indicadas por tu docente en línea y enviar tu archivo. Cabe recordar que esta actividad tiene una ponderación del 10% de tu evaluación.

Para el envío de tu autorreflexión utiliza la siguiente nomenclatura: BIIN_U3_ATR _XXYZ, donde BIIN corresponde a las siglas de la asignatura, U3 es la unidad de conocimiento, XX son las primeras letras de tu nombre, y la primera letra de tu apellido paterno y Z la primera letra de tu apellido materno

Cierre de la unidad

Los avances en la tecnología de secuenciación han revolucionado muchos aspectos de la biología, hoy en día contamos con bases de datos especializadas y de acceso público, que facilitan el estudio del ADN, ARN y las proteínas.

Después de haber hecho un recorrido por algunas de las herramientas que la bioinformática ofrece te podrás dar cuenta que esta disciplina no solamente es útil para el análisis de secuencias y de estructuras, sino que también ofrece muchas opciones para el diseño de moléculas útiles para la ingeniería bioquímica.

Actualmente gracias a la creación de diversos programas, se puede modelar una proteína de interés y su interacción con otras moléculas para poder inferir la reacción que cataliza una enzima o bien para el diseño de moléculas como fármacos.

Finalmente es importante recalcar que debido a los rápidos avances de la tecnología los programas y bases de datos están en continua actualización y mejoramiento, es por ello que la bioinformática ofrece día a día nuevas y mejores metodologías para resolver problemas de la ciencia.

Fuentes de consulta

- Acharya K. R., Stuart D. I., Walker N. P., Lewis M., Phillips D. C. (1989). *Refined structure of baboon alpha-lactalbumin at 1.7 A resolution. Comparison with C-type lysozyme.* J.Mol. Biol. 208:99-127.

- Borden K.L.B., Freemont P.S (1996). *The RING finger domain: a recent example of a sequence-structure family.* Curr. Opin. Struct. Biol. 6:395-401.

- Brown S.M. (2000). *Bioinformatics. A biologist´s guide to biocomputing and the internet.* New York: Eaton Publishing.

- Bru C, Courcelle E, Carrère S, Beausse Y, Dalmar S, Kahn D. (2005) *The ProDom database of protein domain families: more emphasis on 3D*. Nucleic Acids Res.1;33:D212-5.

- Buehler L.K., Rashidi H.H. (Eds.) (2005). *Bioinformatics Basics. Applications in Biological Science and Medicine.* Second Edition. EUA: Taylor & Francis.

- Chis L., Hriscu M., Bica A., Tosa M, Nagy G., Róna G., Vértessy B.G., Irimie F.D. (2013). *Molecular cloning and characterization of a thermostable esterase/lipase produced by a novel Anoxybacillus flavithermus strain.* J. Gen. Appl. Microbiol., 59:119-134.

- Chou KC, Elrod DW. (1999). *Prediction of membrane protein types and subcellular locations.* Proteins. 34(1):137–153.

- Cid H, Bunster M, Canales M, Gazitua F. (1992) *Hydrophobicity and structural classes in proteins.* Protein Eng. 5(5):373–375.

- Combet C., Blanchet C., Geourjon C. Deléage G. (200). *Network Protein Sequence Analysis.* TIBS, 25:[291]:147-150.

- Dokholyana NV, Mirnya LA, Shakhnovicha EI. (2002). *Understanding conserved amino acids in proteins*. *Physica* A, 314:600-606.

- Eisenhaber F, Imperiale F, Argos P, Frommel C. (1996). *Prediction of secondary structural content of proteins from their amino acid composition alone. I. New analytic vector decomposition methods. Proteins.* 25(2):157–168.

- Gromiha MM. (2010*). Protein Bioinformatics: From Sequence to Function.* ELSEVIER. SBN: 978-8-1312-2297-3

- Hartl D. L., Jones E.W. (1998). *Genetics. Principles and Analysis (4º Edition).* EUA: Jones and Bartlett Publishers.

- Hernández-Santoyo A, Tenorio-Barajas AY, Altuzar V, Vivanco-Cid H, Mendoza-Barrera C. (2013). *Protein-Protein and Protein-Ligand Docking, Protein Engineering - Technology and Application*, Dr. Tomohisa Ogawa (Ed.), ISBN: 978-953-51-1138-2, InTech, DOI: 10.5772/56376.

- Jiménez García L. F., Merchant Larios H. (2003). *Biología celular y molecular.* México Pearson Education.

- Koning J., Wanjun G., Castoe T., Batzer M., Pollock D. (2011). *Repetitive elements may comprise over two-thirds of the human genome.* PLoS Genet 7(12): e1002384. doi:10.1371/journal.pgen.1002384

- Medzihradszky KF (2008). *Characterization of site-specific N-Glycosylation post-translational odifications of proteins.* Methods in Molecular Biology, 446:293-316.

- Mount D.W. (2001). *Bioinformatics Sequence and Genome Analysis.* New York: CSHL Press.

- Orengo C.A., Jones D.T., Thornton J.M. (2003). *Bioinformatics: Genes, proteins & computers.* New York: BIOS Scientific Publishers.

- Palczewski K. (2006) *G Protein–coupled receptor Rhodopsin*. *Annu Rev Biochem*. 75: 743–767.

- Pautsch A, Schulz GE. (2000). *High-resolution structure of the OmpA membrane domain.* J Mol Biol. 298:273–282.

- Perez A. (2011*). Predicting Protein Complex Structures: A Review of the Docking Process.* BIOC218 Final Project. Disponible en: http://biochem218.stanford.edu/Projects%202011/Perez%202011.pdf

- Qasba PK, Kumar S (1997). Molecular divergence of lysozymes and alpha lactalbumin. *Crit. Rev. Biochem. Mol. Biol.* 32 (4): 255–306.

- Ratledge C., Kristiansen B (Eds.) (2006). *Basic Biotechnology.* Third Edition. New York: Cambridge University Press.

- Reece R.J. (2004). *Analysis of genes and genomes.* UK: John Wiley & Sons, Ldt.

- Sander C, Schneider R. (1991). *Database of homology-derived protein structures and the structural meaning of sequence alignment.* Proteins. 9(1):56-68.

- Sambrook J., Russell D. (2001). *Molecular cloning: A laboratory manual.* Cold spring harbor, tercera edición. Nueva York, USA. Segundo volumen, sección 8.13.

- Schuler, G. Sequence Alignment and Database searching en: Baxevanis A.F., Ouellette, F. (2001). *Bionformatics: A practical guide to the analysis of genes and proteins*. 3ª Ed. UK: Wiley.

- Shin D.H., Roberts A., Jancarik J., Yokota H., Kim R., David E. W. and Kim S. H. (2003). *Crystal structure of a phosphatase with a unique substrate-binding domain from Thermotoga marítima.* Protein Science, 12:1464–1472.

- Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I. (2012). *New and continuing developments at PROSITE.* Nucleic Acids Res; doi: 10.1093/nar/gks1067.

- Smith GR y Sternberg MJE (2002). *Prediction of protein–protein interactions by docking methods Current Opinion in Structural Biology*. 12:28–35.

- Twyman RM. (2004). *Principles of proteomics*. New York. BIOS Scientific Publishers.

- von Heijne G. (1986). *The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology*. *EMBO J*. 5:3021–3027.

- Wallin E, von Heijne G. (1998). *Genome-wide analysis of integral membrane proteins from*
- *eubacterial, archaean, and eukaryotic organisms*. *Protein Sci.* 7(4):1029–1038.

- Westhead D.R., Parish J.H., Twyman R.M. (2002). *Instant Notes Bioinformatics.* New York. BIOS Scientific Publishers.

- Zhang Z, Sun ZR, Zhang CT. (2001). *A new approach to predict the helix/strand content of globular proteins. J Theor Biol*. 208:65–78.