PARTE I: Predicción de Desorden

Lucía B. Chemes, Juliana Glavina

Recursos a utilizar:

- ProViz http://slim.icr.ac.uk/proviz/
- DisProt https://www.disprot.org
- IUPred2A https://iupred2a.elte.hu/plot
- MobiDB http://mobidb.bio.unipd.it/

Introducción

Objetivos.

- Interpretar alineamientos múltiples de secuencias
- Familiarizarse con distintos métodos de predicción de desorden
- Interpretación de los resultados de los distintos métodos

Métodos de predicción de desorden

Uno de los mayores desafíos en el campo de las proteínas es la predicción de la estructura tridimensional a partir de la estructura primaria incluyendo aquellas proteínas que son total o parcialmente desordenadas. Mientras que las proteínas globulares adquieren una única estructura nativa, las proteínas intrínsecamente desordenadas (IDPs) son un conjunto de estructuras tridimensionales. También pueden existir regiones de proteínas que pueden ser desordenadas como por ejemplo fragmentos proteicos que conectan dos dominios globulares, denominados *loops* o regiones que abarcan más de 30 residuos de longitud en cuyo caso se los llama regiones intrínsecamente desordenadas (IDRs).

La predicción de IDRs a partir de la secuencia de aminoácidos permite un análisis rápido y abarcativo de distintas proteínas permitiendo establecer hipótesis sobre la presencia de desorden en las proteínas (Dunker et al., 2008; van der Lee et al., 2014). La importancia que adquirieron las IDRs/IDPs en los últimos años llevó al desarrollo de numerosos métodos de predicción, pero en general se basan en tres estrategias de predicción de desorden: (1) a partir de composición de secuencia, (2) a partir de *machine learning* sobre estructuras determinadas por cristalografía de rayos X y (3) a partir de meta-predictores que integran los resultados predichos por diferentes métodos.

Entre los algoritmos que se basan en composición de secuencia podemos nombrar IUPred (Dosztányi et al., 2005a,b; Mészáros et al., 2018), que aplica un campo de energía desarrollado a partir de un gran número de proteínas con estructura determinada obtenidas de PDB. El primer algoritmo en *machine learning* fue PONDR (Obradovic et al., 2003; Romero et al., 1997), entrenado a partir de un

grupo estructuras de proteínas globulares y atributos de secuencia asociados a residuos no resueltos en dichas estructuras, que corresponden a regiones flexibles dentro del cristal. GlobPlot (Linding et al., 2003b) fue entrenado estudiando la tendencia de un residuo a adquirir determinada estructura secundaria, hélices α o láminas β.

Guía de Ejercicios - Desorden

Ejercicio 1. Visualización de Alineamientos en ProViz.

Antes de empezar, piensen: ¿Porqué es importante visualizar un MSA? ¿Qué información podemos obtener de los MSA?

ProViz es una herramienta que permite visualizar alineamientos y estructura de dominios de una proteína. Ingresa a la web de proviz(http://slim.icr.ac.uk/proviz/), y busca la proteína Calcineurina A ingresando su UNIPROT ID en la ventana "search" (UNIPROT ID: Q08209):

Selecciona la proteína que se llama: Serine/threonine-protein phosphatase 2B catalytic subunit alpha isoform (PPP3CA) Homo sapiens (Human). Es la primera de la lista. Chequea que en la parte superior de la página, en **alignments** esté seleccionada la opción **QFO**.

1. ¿Qué regiones parecen estar mejor alineadas?
2. ¿Existe diferencia en la composición de secuencia?
3. ¿Se observan diferencias en el grado de conservación?
4. ¿A qué pueden deberse las diferencias observadas?

Ejercicio 2. Predicción de desorden a partir de secuencia.

Ingresa en la web de IUPred2A (https://iupred2a.elte.hu) e ingresa la proteína Calcineurina A (puede ingresarse la secuencia de aminoácidos, el UNIPROT ID, PP2BA_HUMAN, o el accession number, Q08209). El algoritmo IUPred considera que un residuo es desordenado cuando el valor de IUPred es mayor o igual a 0.5 y ordenado cuando es menor a 0.5. Anota las posiciones iniciales y finales de las regiones predichas como desordenadas. ¿Se correlacionan las regiones predichas como ordenadas o desordenadas con las diferencias observadas en el ejercicio anterior?

Ejercicio 3. Base de datos DisProt

Objetivos:

- Familiarizarse con la base de datos DisProt
- Entender las técnicas experimentales que permiten la identificación de regiones desordenadas.

Introducción a Disprot: La base de datos DisProt es una colección de evidencia de desorden experimental recolectada de la literatura y curada manualmente. La evidencia corresponde a una región proteica, e incluye por lo menos: un experimento, el artículo científico correspondiente a ese experimento, el inicio y final de la región en la secuencia proteica y un término de anotación que

corresponde a la Ontología de desorden. ***Cada una de las entradas en la base de datos posee un identificador único***.

La ontología de desorden está organizada en cinco categorías diferentes:

1. Estado estructural (*Structural State*): Order or Disorder
2. Transición estructural (*Structural Transition*): Transiciones que pueden ocurrir entre diferentes estados estructurales (Disorder to order)
3. Par de Interacción (*Interaction Partner*): La entidad que interactúa (proteína, ión, moléculas pequeñas)
4. Función de desorden (*Disorder Function*): La función de una región incluyendo términos específicos a desorden.
5. Método experimental (*Experimental Method*): Métodos experimentales para detectar regiones desordenadas.

Ejercicio:

La proteína Calcineurina A es una proteín fosfatasa estimulada por calmodulina calcio dependiente. Posee un rol importante en la transducción de las señales intracelulares mediadas por Ca^{2+}. En respuesta a los aumentos de Ca^{2+}, Calcineurina A desfosforila diversas proteínas. Por ejemplo, desfosforila y activa la fosfatasa SSH1 que lleva a la desfosforilación de Cofilina (una proteína de unión a actina que desensambla los filamentos de actina).

1. Ingresa a la página web de DisProt (www.disprot.org) y encuentra la proteína Calcineurina A (PP2BA_HUMAN, Q08209). La búsqueda puede realizarse utilizando el Accession Number o por palabras claves. El identificador de DisProt que deberían encontrar es DP00092.
2. Expande "*Disprot consensus*" ¿Qué tipo de información observa en la página?
    a. Expande *"Structural state"* y luego expande *"Disorder"*. ¿A qué corresponden los segmentos coloreados? ¿Qué tipo de evidencia poseen dichos fragmentos?
3. ¿Cuál es el rol de las regiones desordenadas?
    a. Expande "Interaction" ¿Qué tipo de interacciones están indicadas?¿Qué técnicas se usaron para identificarlas?
    b. Expande "Function" ¿Qué tipo de funciones están indicadas? ¿Qué técnicas se usaron para identificarlas?
4. ¿Se observa algún dominio globular conservado?
    a. Expande "Domains". ¿A qué corresponden los segmentos coloreados? ¿Qué tipo de evidencia poseen dichos fragmentos?
5. ¿La evidencia experimental recolectada coincide con las predicciones realizadas en los ejercicios 1 y 2?

Ejercicio 4. Selección de regiones para determinar la estructura de una proteína.

Una de las aplicación principales de la predicción de desorden es encontrar regiones que son más adecuadas para determinar la estructura tridimensional de una proteína por cristalografía de rayos X.

1. ¿Por qué cree que predecir las regiones desordenadas puede ayudar a seleccionar el dominio para cristalizar?

Dada la siguiente proteína misteriosa:

```
>mystery_protein
MMQDLRLILIIVGAIAIIALLVHGFWTSRKERSSMFRDRPLKRMKSKRDDDSYDEDVEDDEGVGEVRVHRVNH
APANAQEHEAARPSPQHQYQPPYASAQPRQPVQQPPEAQVPPQHAPHPAQPVQQPAYQPQPEQPLQQPVSPQV
APAPQPVHSAPQPAQQAFQPAEPVAAPQPEPVAEPAPVMDKPKRKEAVIIMNVAAHHGSELNGELLLNSIQQA
GFIFGDMNIYHRHLSPDGSGPALFSLANMVKPGTFDPEMKDFTTPGVTIFMQVPSYGDELQNFKLMLQSAQHI
ADEVGGVVLDDQRRMMTPQKLREYQDIIREVKDANA
```

2. Utilizando IUPred2A, pega solamente la secuencia sin el header ¿Qué región de la proteína trataría de cristalizar?
3. Para ver si la selección fue la correcta, haz un blast de la secuencia en la página web https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome.

   Pega la secuencia en el box “Enter Query Sequence”. Chequear que el box “align two or more sequences” no esté seleccionado.

   En la sección **Choose Search Set**, selecciona la **database Protein Data Bank proteins (pdb)**.

   Explora los resultados. ¿Elegimos correctamente?

Nota: El predictor de desorden DisMeta cuya página web es:

http://www-nmr.cabm.rutgers.edu/bioinformatics/disorder/

se desarrolló específicamente para diseñar construcciones para cristalografía de rayos X. El método es muy lento. Por lo tanto, explora los resultados en casa.

PARTE II - Motivos Lineales en Proteínas

Recursos a utilizar:

- Regex101 https://regex101.com
- UniProt http://www.uniprot.org/

- ELM http://elm.eu.org

Ejercicio 1. Familiarizándonos con las Expresiones Regulares

Objetivos:

- Familiarizarse con la simbología utilizada en expresiones regulares
- Utilizar la simbología para poder realizar búsquedas basadas en texto

La simbología comúnmente utilizada en expresiones regulares es:

Símbolo	Definición
`.`	Cualquier aminoácido es permitido
`[XY]`	Solo los aminoácidos X e Y son permitidos
`[^XY]`	Los aminoácidos X e Y están prohibidos
`{min,max}`	Número mínimo y máximo de veces que se puede repetir una posición
`^X`	El aminoácido X se encuentra en el extremo N-terminal
`X$`	El aminoácido X se encuentra en el extremo C-terminal
`(AB)|(CD)`	Se encuentran, o bien, los aminoácidos AB, o bien, los aminoácidos CD

Estos símbolos nos permiten definir patrones que son observados en proteínas naturales para luego identificarlos en otras proteínas y ser puestos a prueba experimentalmente.

Los receptores nucleares interactúan con diversas proteínas mediantes un motivo lineal llamado NRBox (*Nuclear Receptor Box*) (Heery,1997). Existen numerosas estructuras de péptidos unidos a diferentes receptores nucleares (PDBs: 3CS8, 2GPO, 1GWQ, 1RJK, 1M2Z) que permitieron estudiar y entender algunas características de la interacción.
La evidencia experimental recolectada de la literatura indica que:

I. El motivo NRBox forma una hélice alfa
II. Existen tres leucinas que se **encuentran en una misma cara de la hélice** que interactúan con un bolsillo hidrofóbico en la superficie del receptor nuclear (Figura 1).

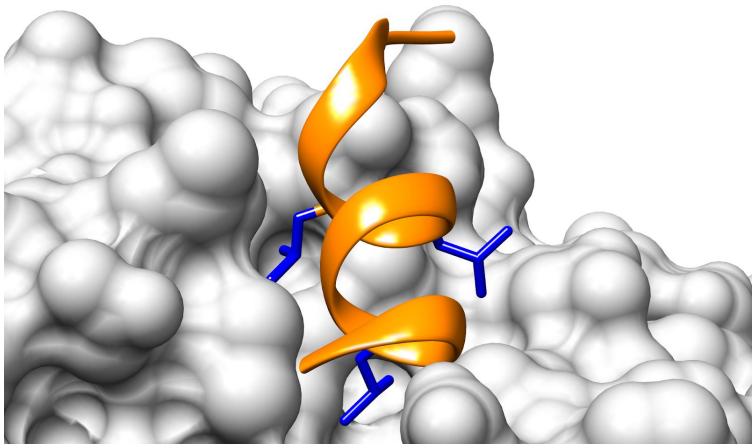


Figura 1. Fragmento de la proteína PGC-1 alfa unido al receptor nuclear PPAR-gamma. Se muestra en naranja el backbone de la proteína representado en “Cartoon” y en azul las tres leucinas que median la interacción representadas en “Sticks” (PDB:3CS8) y que conforman el motivo NRBox.

Los siguientes fragmentos de secuencia corresponden a regiones de distintas proteínas que interactúan con diversos receptores nucleares y cuya interacción se verificó de manera experimental por distintos métodos.

```
>sp|Q15648|MED1_HUMAN|644-650
SMAGNTKNHPMLMNLLKDNPAQDFSTL
>sp|O43593|HAIR_HUMAN|565-571
AKHLLSGLGDRLCRLLRREREALAWAQ
>sp|Q16881-4|TRXR1_HUMAN|46-52
GPTLKAYQEGRLQKLLKMNGPEDLPKS
>sp|P48552|NRIP1_HUMAN|500-506
DVHQDSIVLTYLEGLLMHQAAGGSGTA
>sp|Q9UQ80|PA2G4_HUMAN|353-359
YKSEMEVQDAELKALLQSSASRKTQKK
>sp|Q90ZL7|Q90ZL7_DANRE|69-75
VQHADGEKSNVLRKLLKRANSYEDAVM
>sp|Q9UBK2|PRGC1_HUMAN|143-149
PPPQEAEEPSLLKKLLLAPANTQLSYN
>sp|Q9JL19|NCOA6_MOUSE|1494-1500
MSPAMREAPTSLSQLLDNSGAPNVTIK
>sp|Q15596|NCOA2_HUMAN|689-695
HGTSLKEKHKILHRLLQDSSSPVDLAK
>sp|Q92793|CBP_HUMAN|69-75
LVPDAASKHKQLSELLRGGSGSSINPG
```

a. Copia y pega las secuencias en el recuadro de *Test String* en regex101 y prueba encontrar una expresión regular que permita identificar el motivo que media la interacción de estas proteínas con los receptores nucleares y que cumpla con la evidencia experimental observada.

b. Considerando que el motivo se encuentra en una hélice, ¿modificaría la expresión regular que obtuvo?
c. Busca en ELM (http://elm.eu.org) en la pestaña ***Prediction*** una de las proteínas de la lista que usamos recién: la proteína PGC-1-alpha utilizando el accession number o uniprot ID (Q9UBK2 - PRGC1_HUMAN). Para cada motivo encontrado, se indica con símbolos (descriptos en la parte superior de la página) si la instancia del motivo es predicha o fue identificada experimentalmente (anotadas o "True Positives"). ¿Encuentra el motivo **NRBox** entre los true positives? ¿Cuantas instancias "True Positive" existen? ¿Cómo es la estructura de la proteína donde se encuentran estos motivos?

PARTE III: Análisis de alineamientos múltiples de secuencia de proteínas - Visualizando alineamientos con JalView

Recursos a utilizar:

- JalView: https://www.jalview.org/
- PFAM: https://pfam.xfam.org/

Introducción

Objetivos:

- Familiarizarse con el manejo de programas de visualización de alineamientos.
- Interpretar alineamientos múltiples de secuencias.
- Familiarizarse con distintas visualizaciones del alineamiento.

JalView, software de visualización de alineamientos.

Para poder visualizar alineamientos múltiples de secuencias (MSA, de sus siglas en inglés: Multiple Sequence Alignment) utilizaremos el visualizador de alineamientos JalView desarrollado en JAVA. Jalview permite generar alineamientos, manipularlos, editarlos y anotarlos. Tiene una interfaz que permite acceder remotamente numerosas herramientas como programas para realizar alineamientos múltiples de secuencia y predictores de estructura secundaria. A lo largo de la guía de ejercicios, introduciremos este programa usandolo para visualizar alineamientos múltiples de secuencias (MSAs) de proteínas modulares y discutir características de secuencia asociadas a los dominios y motivos funcionales encontrados en las proteínas.

JalView es un programa disponible de manera gratuita, y está disponible para descargar e instalar en tu propia computadora en https://www.jalview.org/

Existen un alto número de guías y tutoriales disponibles online que pueden encontrarse en: https://www.jalview.org/training
Los desarrolladores de JalView crearon numerosos videos de entrenamiento disponibles en el Canal de YouTube de JalView

Guía de Ejercicios - JalView

Objetivos:

- Aprender a utilizar Jalview para visualizar un MSA
- Identificar regiones de secuencia conservadas y asociarlas a diferentes elementos funcionales de las proteínas.
- Visualizar y analizar los patrones de sustitución aminoacídica encontrados en proteínas modulares. Correlacionar con sus conocimientos sobre matrices de sustitución

Ejercicio 1. Identificando Módulos en Proteínas

Utilizando su código UNIPROT (P04637), busca la proteína p53 humana (P53_HUMAN) en la base de datos PFAM. https://pfam.xfam.org/

La base de datos PFAM es una colección de familias de dominios de proteínas construída en base a alineamientos múltiples de secuencia y modelos ocultos de markov (HMMs). Las proteínas están compuestas por una o más regiones funcionales o dominios, que combinados de distintas maneras crean la diversidad proteica que se encuentra en las proteínas naturales.

¿Porqué es necesario identificar dominios en las proteínas?

Para buscar la proteína p53 puedes hacerlo ingresando en VIEW A SEQUENCE el accession number (P04637) o el uniprot ID (P53_HUMAN)

1. ¿Qué longitud tiene la proteína p53 humana?
2. Observar el esquema modular de p53: ¿Puedes identificar qué dominios PFAM tiene p53? ¿Qué nombres y qué funciones tienen?
3. ¿En qué regiones de la secuencia se encuentran estos dominios? Anotar de qué residuo a qué residuo abarca cada dominio, para usar más adelante.
4. ¿Creen que estos dominios corresponden unívocamente a dominios globulares?
5. ¿A qué cree que corresponden las regiones marcadas como “Disorder” y “Low Complexity” en p53?
6. Ingresa la proteína en IUPred. ¿Se corresponden las regiones identificadas como Disorder en PFAM con las predichas por IUPred?

Ejercicio 2. Usando JalView para analizar un MSA de p53

La proteína p53 es una proteína supresora de tumores, es decir que su mutación favorece el crecimiento tumoral. p53 es uno de los genes más mutados en el cáncer humano, y actúa como un

factor de transcripción que se expresa en todos los tejidos. Cumple un rol principal en el ciclo celular y es el regulador principal de la apoptosis. Es esencial para inducir la respuesta celular ante el daño al ADN, deteniendo el ciclo celular cuando las células no pueden reparar el ADN dañado por agentes genotóxicos. Si falla p53 podrían facilitar la formación de tumores celulares y en consecuencia producir cáncer. Alrededor de un 50% de los tumores humanos identificados poseen mutaciones en la proteína p53. Esta proteína, por su importancia para la salud humana, es una de las proteínas más estudiadas en cuanto a su estructura y función.

1. Descarga un conjunto de secuencias homólogas de p53 obtenido de la base de datos Swiss Prot. El archivo también se encuentra en la carpeta MSA del TP de la materia y se llama p53.fasta

   File → Input Alignment → From File

2. Para realizar el alineamiento utilizaremos el programa Clustal, al cual accederemos de manera remota desde JalView:

   Web Service → Alignment → Clustal → With defaults

   (O descarga y abre el archivo **p53_aligned.fasta** que se encuentra en la carpeta MSA del TP de la materia)

3. Inspecciona el alineamiento visualmente y reconoce algunas características de las secuencias. Si no se muestran todos los residuos y algunos aparecen como "." ve a:

   Format → Show Non-Conserved

   a. Algunas secuencias son más cortas que otras ¿por qué crees que es esto?
   b. ¿Todas las secuencias comienzan con el aminoácido metionina? A qué corresponden las secuencias que no?
   c. ¿Si quieren construir un alineamiento de alta calidad, preservarían o descartarían estas secuencias?
   d. Remuevan las secuencias que no corresponden a proteínas completas. Para ello seleccionar las secuencias haciendo click sobre el nombre de la misma en el panel izquierdo, la secuencia se marcará con una caja roja punteada. Remover la secuencia seleccionada utilizando la tecla "*Backspace*" o "*Del*"
   e. ¿Existen regiones del alineamiento que no estén alineadas correctamente?
      Para editar el alineamiento, primero asegurate de realizar:

      Select → Deselect All

      Eliminar gaps: Seleccione con el mouse el gap o arrastrando sobre el grupo de gaps que desea eliminar y presione "Backspace" o "Del"

      Agregar gaps: Presione F2. En primera posición del alineamiento en la primera secuencia aparecerá un cursor de color negro. Colóquelo en la posición donde desee ingresar un gap y presione la barra espaciadora.

Ejercicio 3. Análisis de distintas propiedades del MSA utilizando el menú *COLOUR*.

Este menú permite colorear el alineamiento con diferentes paletas de colores que permiten visualizar determinadas características fisicoquímicas o relacionadas con la conservación o identidad de secuencia que facilitan el análisis de la información contenida en el MSA.
Por ejemplo: *Percentage identity* colorea los residuos según el porcentaje de identidad en la columna. *Hydrophobicity* colorea los residuos según el grado de hidrofobicidad.
También es posible disminuir la intensidad de los colores según el grado de conservación (*By conservation*) o filtrar los colores según el porcentaje de identidad (*Above identity threshold*) a partir de un umbral deseado.

1. Seleccione para colorear el alineamiento desde el menú la opción:

   Colour → ClustalX

   Este esquema es muy comúnmente utilizado para la visualización de MSAs y permite representar información importante contenida en los patrones de sustitución de un MSA

   Observando el alineamiento intente identificar:
   a. ¿Cuál es la base del esquema de color “ClustalX” provisto por Jalview?
   b. ¿Cuantos colores existen?
   c. ¿Qué propiedades fisicoquímicas representa cada grupo de color?
   d. La cisteína cumple un rol estructural importante en algunas proteínas (cual?). Qué observa respecto de la coloración de la cisteína: ¿Es siempre igual? ¿a qué se debe el cambio en la representación?
   e. ¿En qué situaciones los residuos no están coloreados?
   f. Hay residuos que siempre están coloreados? Cuales son y a qué cree que se debe?

2. Manteniendo el esquema de color Clustal, es posible filtrar regiones de acuerdo al % identidad en el alineamiento múltiple. Para ello, aplique el filtro de identidad yendo a:

   Colour → Above identity threshold

   Se abrirá una ventana en la cual podrá seleccionar el % identidad del filtro en escala de 0 a 100%. Explore los cambios en todo el alineamiento al variar la escala de 0 a 100%.

   Utilizando el filtro, respondan:

   a. ¿Qué regiones muestran una identidad de secuencia mayor al 80% en el MSA de p53? ¿Y al 100%? Anote los límites de estas regiones y responda: ¿Qué correlación observa con la información obtenida de PFAM?

   En las regiones conservadas, observe los patrones de sustitución en diferentes columnas del MSA. Estos patrones son un reflejo de la historia evolutiva de la proteína y contienen mucha información funcional que aprenderemos a cuantificar más adelante en la materia. Observando detenidamente, responda:

b. ¿Qué tipos de patrones observa?
c. ¿Qué relación guardan estos patrones con las matrices PAM y BLOSUM utilizadas para construir alineamientos de proteínas?

Ejercicio 4. Identificación de motivos cortos de interacción en p53.

La región amino terminal de p53 posee un motivo de unión a la E3 ligasa MDM2, el cual está caracterizado por una secuencia conservada que puede representarse por una expresión regular.

1. i- Entra en la base de datos ELM y busca la expresión regular del motivo con el ID: DEG_MDM2_SWIB_1. Para esto ingresa el ID en la parte superior derecha donde dice: *Search ELM database. La expresión regular se encuentra marcada como "Pattern"*

   ii- A continuación, busca las ocurrencias de esta expresión regular en las secuencias de p53. Para ello, Jalview permite la búsqueda de motivos por expresiones regulares. Para hacerlo, utilice la función:

   Select → Find

   En la ventana tipee la expresión regular. Si este procedimiento falla, asegúrese de tener la ventana de las secuencias no alineadas cerrada. Si aún así falla, identifique el motivo utilizando el filtro de conservación

   a. ¿Todas las secuencias de p53 tienen el motivo de interacción con MDM2?
   b. ¿Todos los motivos MDM2 tienen la misma longitud?
   c. ¿Qué nivel de identidad de secuencia observa en esta región? ¿A qué puede deberse?

Ejercicios Adicionales

Desorden

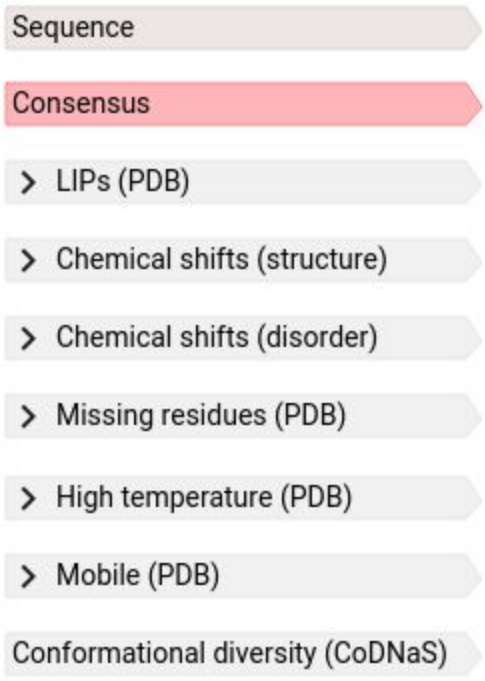
Ejercicio Adicional 1. Base de datos MobiDB

La base de datos MobiDB centraliza diferentes recursos que facilitan la anotación de proteínas desordenadas y de su función. MobiDB abarca distintos aspectos del desorden, desde regiones que carecen una estructura tridimensional definida anotadas o predichas como desordenadas hasta regiones que interactúan con otras proteínas, ADN o ARN preservando una estructura desordenada. Los datos provienen de bases de datos externas con datos manualmente curados, de datos experimentales como estructuras tridimensionales de las proteínas o predicciones.

1. Ingresa a la web de MobiDB (http://mobidb.bio.unipd.it) y busca la proteína Calcineurina A (Q08209).

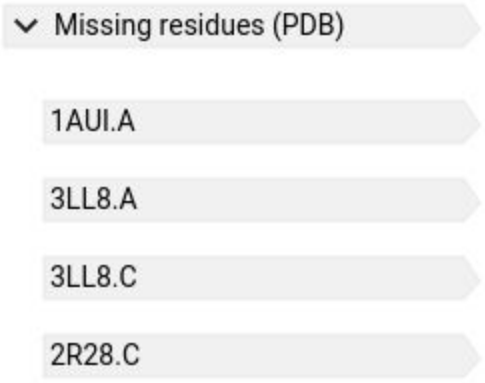
2. Ve a la pestaña *Predictions*. ¿Cuáles regiones son predichas como desordenadas por la mayoría de los métodos? ¿Qué métodos predicen más desorden y cuáles menos? ¿Hay mucha variación?

3. Ve a la pestaña *Indirect*. A la izquierda se ve lo siguiente:



   En la primera línea se indica la secuencia y en la segunda línea (*Consensus*) se indica el consenso en base a la evidencia estructural. Ubique el mouse sobre las distintas regiones y responda: ¿Qué significan los distintos colores de las regiones marcados en el consenso?

4. Exploremos la evidencia proveniente de estructura cristalográfica. Para eso despliegue la sección *Missing residues* (PDB).



   ¿Qué regiones tienen una estructura?

Ve a la primera entrada (1AUI_A) y cliquea en el último botón de la línea (*"go to PDB"*). En la web de la base de datos de PDB ve a la sección *Macromolecules*. Mira la sección correspondiente a la cadena A. ¿Puedes decir cómo se determinó que estas regiones eran desordenadas?

5. Vuelva a la pestaña de MobiDB Existen regiones de la proteína que presentan evidencia conflictiva de desorden en el consenso (marcadas como *conflict*). Mirando las distintas estructuras resueltas en MobiDB responda ¿Por qué estas regiones están marcadas como conflictivas?

Ejercicio Adicional 2. Búsqueda de regiones funcionales dentro de las IDPs, usando como ejemplo la proteína p53.

Objetivos:

- Familiarizarse con la identificación de sitios de unión en IDPs
- Interpretación de los resultados de los distintos métodos.

Muchas proteínas desordenadas ejercen su función uniéndose a una proteína globular, mediante una transición de desorden a orden. ANCHOR es un algoritmo para predecir sitios de unión en proteínas desordenadas buscando identificar segmentos que residen en regiones desordenadas y no forman interacciones intracatenarias suficientes que favorezcan el plegado por sí mismas, pero si logran estabilizarse al interactuar con una proteína globular.

1. Ve a la web de IUPred. https://iupred2a.elte.hu

2. Ingresa la proteína p53 (P53_HUMAN), asegúrate que la opción ANCHOR en "Context-dependent predictions" esté seleccionada.

   ¿Cuántas regiones de interacción identifica ANCHOR?

3. La base de datos IDEAL se enfoca en IDRs que adoptan una estructura 3D al unirse a sus pares proteicos y se los llama *Protean Segments* (ProS), que se definen cuando la información estructural y no *desestructural* existen. Hay otros conceptos similares a los ProS que difieren en la definición, como por ejemplo, los *Molecular recognition features (MoRFs)*, que tienen una limitación de longitud de 70 residues y los motivos lineales eucarióticos que son expresados por expresiones regulares.

   Ingresa a la base de datos IDEAL y busca la proteína p53 (P53_HUMAN, P04637). ¿Qué regiones están involucradas en la formación de complejos?

   Prestando atención a la región C-terminal:

   a. ¿A cuántas proteínas distintas se une p53?

   b. ¿Qué tipo de estructura secundaria adquieren en el complejo?

4. Busca los PDBs: 1MA3, 1H26, 1JSP, 1DT7.

   ¿Cuán parecidas son las predicciones de ANCHOR con las regiones de unión conocidas?

NOTA: Existen muchísimos métodos para predecir regiones desordenadas. Puedes probar los siguientes métodos en casa y ver las diferencias:

- PONDR http://www.pondr.com
- PredictProtein http://ppopen.informatik.tu-muenchen.de/ (IDPs se predicen por Meta-Disorder a partir de una combinación de NORSnet, DISOPRED2, PROFbval y Ucon)
- Globplot2 http://globplot.embl.de/
- DISOPRED3 http://bioinf.cs.ucl.ac.uk/psipred/ (Elegir la opción Disopred3). Este método lleva por lo menos 20 minutos y puede tardar hasta 2 horas.

Ejercicio Adicional 3. Análisis de una proteína altamente desordenada

1. Utiliza un predictor de desorden para la entrada de DisProt DP00039
2. Utiliza el servidor protparam (https://web.expasy.org/protparam/), o algún otro método que conozcas, para contar el número de aminoácidos cargados positivamente y el número de aminoácidos cargados negativamente.
3. Calcula la carga neta (o utiliza el servidor protparam)
4. Observa los segmentos de baja complejidad de secuencia (indicados en PFAM)
5. Observa los dominios PFAM.
6. ¿Existen contradicciones entre la asignación de dominios PFAM y el desorden predicho?

Ejercicio Adicional 4. Caracterización de la proteína humana N-WASP (O00401) desde el punto de vista de orden y desorden.

1. Busca el número de estructuras PDB que existen para esta proteína (http://www.rcsb.org/pdb/protein/O00401 → “Number of PDB entries for O00401”
2. ¿Qué regiones de la proteína N-WASP están resueltas para cada entrada del PDB?
3. Busca familias PFAM y observa el tipo.
   a. Haz click en el domain
   b. Haz click en “Curation and model”
   c. Chequea el tipo: “Domain”, “Family” o “Motif”
4. Encuentra regiones de baja complejidad (“low complexity”) ¿Qué aminoácidos son más frecuentes en esta región?
5. Utiliza el predictor de desorden de tu preferencia.
6. ¿Qué regiones llamarías desordenadas?

Motivos Lineales

Ejercicio adicional 1. Un poquito más de expresiones regulares

La reparación del ADN durante la replicación ocurre por un proceso llamado Translesion synthesis (TLS). En este proceso, una polimerasa TLS, inserta un nucleótido en la lesión del ADN y luego, una polimerasa de la familia B extiende el templado. La acción coordinada de estas polimerasas, se logra por la interacción de proteínas scaffold como PCNA (*Proliferating Cell Nuclear Antigen*) y la polimerasa TLS Rev1.

Existen estructuras cristalográficas de distintos péptidos unidos a Rev1 (PDBs: 2N1G, 2LSK, 2LSJ, 4FJO, 2LSI y 4GK5) que permiten entender algunas características de la interacción.

La evidencia experimental recolectada de la literatura indica:

I. La interacción está mediada principalmente por dos residuos consecutivos de fenilalanina (Ohashi,2009).
II. Las fenilaninas interactúan con un bolsillo hidrofóbico en la superficie de Rev1 (Pozhidaeva, 2012; Zhao,2017).
III. Las fenilalaninas se encuentran en el primer giro de una hélice α.
IV. Se requieren al menos 4 residuos posteriores a las fenilalaninas que formen parte de una hélice (Ohashi, 2009)
V. El resto de la región de interacción se pliega formando hélices α de longitud variable (Pustovalova, 2016)
VI. En general se observan residuos cargados positivamente en la 2da y/o 3ra posición luego de las fenilalaninas que median interacciones electrostáticas con una superficie acídica de Rev1. Aunque la posición de estos residuos puede variar.

Los siguientes fragmentos de secuencia corresponden a regiones de distintas proteínas que participan en la reparación del ADN que se unen la proteína Rev1 y cuya interacción se verificó de manera experimental por distintos métodos.

```
>sp|Q03834|MSH6_YEAST|31-38
SQKKMKQSSLLSFFSKQVPSGTPSKKVQ
>sp|Q04049|POLH_YEAST|625-632
KKQVTSSKNILSFFTRKK
>sp|Q60596|XRCC1_MOUSE|191-200
DDSANSLKPGALFFSRINKTSSASTSDPAG
>sp|Q9H040|SPRTN_HUMAN|418-428
RPRLEDKTVFDNFFIKKEQIKSSGNDPKYST
>sp|Q15054|DPOD3_HUMAN|236-245
NKAPGKGNMMSNFFGKAAMNKFKVNLDSEQ
>sp|Q9UNA4|POLI_HUMAN|569-579
SCPLHASRGVLSFFSKKQMQDIPINPRDHLS
```

```
>sp|Q9Y253|POLH_HUMAN|481-490
TATKKATTSLESFFQKAAERQK VK EA SL SS
>sp|Q9Y253|POLH_HUMAN|529-539
PFQTSQSTGTEPFFKQKSLLLKQKQLNNSSV
>sp|Q9QUG2|POLK_MOUSE|564-575
LAKPLEMSHKKSFFDKKRSERISNCQDTSRCK
>sp|Q9UBT6|POLK_HUMAN|565-576
FVKPLEMSHKKSFFDKKRSERKWSHQDTFKCE
```

a. Copia y pega las secuencias en el recuadro de *Test String* en regex101 y prueba encontrar una expresión regular que permita identificar el motivo que media la interacción de estas proteínas con Rev1 y que cumpla con la evidencia experimental observada.
b. Busca en ELM alguna de las proteínas. ¿Tu expresión regular difiere mucho de la propuesta por ELM?
c. SlimSearch es una herramienta que utilizando expresiones regulares permite buscar la presencia de motivos en las proteína almacenadas en Uniprot.
   Ve a la web de SlimSearch (http://slim.ucd.ie/slimsearch/) e ingresa la expresión regular del motivo como figura en ELM.
   i. ¿Cuántas proteínas obtuviste?
   ii. ¿Cuál es la localización celular de Rev1 (Q9UBZ9)? Explora la lista de proteínas. ¿Hay alguna que no tenga la misma localización?
   iii. Encuentra en la lista la proteína Kinesin-like protein KIF11 (P52732). ¿Cuál es su localización?
d. ProViz es una herramienta que colecta y muestra información desde distintas fuentes facilitando la detección de motivos lineales.
   Ingresa en el servidor de Proviz y busca la proteína Kinesin-like protein KIF11 (KIF11) (P52732). Ubica la región donde se encuentra el posible motivo sugerido por SlimSearch. ¿Está conservado? ¿A qué se debe esa conservación? ¿Te parece que es un posible motivo?

Ejercicio Adicional 2. Base de datos de motivos lineales en Eucariotas (ELMdb)

La base de datos ELM (*Eukaryotic Linear Motifs*) es una base de datos que se enfoca principalmente en la anotación y detección de motivos lineales (MLs). Para ello cuenta con un repositorio de motivos manualmente anotados, por lo cual está altamente curada y una herramienta de predicción de motivos. Esta predicción de motivos se realiza mediante una búsqueda de patrones de secuencia basada en texto utilizando expresiones regulares.

Objetivos:

- Familiarizarse con la herramienta de predicción de motivos de ELM.
- Aplicar la herramienta de predicción de motivos de ELM a una proteína viral.

La familia viral *Adenoviridae* (adenovirus) son virus ADNdc desnudos. Los adenovirus que infectan a humanos son responsables de muchas enfermedades respiratorias y de numerosos casos de gastroenteritis en niños. El único género de adenovirus que posee la proteína E1A es el género *Mastadenovirus* que infecta a **mamíferos**. Hasta la fecha, no existe ningún homólogo reportado en

los restantes géneros de esta familia viral. La proteína E1A posee un rol importante en la replicación del genoma viral ya que desregula el ciclo celular induciendo la división celular. Esta estimulación de la progresión de la fase G1 a la fase S, permite que el virus use la maquinaria celular de replicación del ADN para replicar su propio genoma. Una vez expresada la proteína E1A su localización en la célula infectada es **nuclear** y minoritariamente **citosólica**.

1. Busca en ELM (http://elm.eu.org) en la pestaña *Prediction* la proteína E1A del virus Human adenovirus 5 (E1A_ADE05).

a. Utiliza los siguientes parámetros:
   Cell Compartment: **Not specified**
   Motif Probability Cutoff: **100**
   Taxonomic context: **(leave blank)**

   ¿Cuántas clases y cuántas instancias de motivos encuentras?

b. En base a los conocimientos que poseemos de E1A_ADE05 modifica los parámetros Cell Compartment (se puede seleccionar más de un compartimento celular utilizando la tecla ctrl) y taxonomic context. ¿Cómo cambia el número de motivos encontrados?
c. ¿Qué otros filtros observas que está utilizando ELM? ¿Por qué se te ocurre que se eligen automáticamente esos filtros?
d. ¿Qué se puede decir de la estructura de la proteína E1A? ¿Se observa algún dominio? ¿Se observan regiones desordenadas?
e. En cada una de las clases de motivos encontrados, se indica con distintos símbolos (descriptos en la parte superior de la página) si la instancia del motivo es predicha o fue identificada experimentalmente (instancias anotadas o “True Positives”). ¿Cuantas instancias anotadas existen?
f. E1A tiene dos motivos de interacción con la proteína Retinoblastoma, un regulador del ciclo celular (motivo AB_groove y motivo LxCxE). Explora la clase LIG_Rb_LxCxE_1, para esto haz click sobre el nombre de la misma, en la lista de la izquierda. Se abrirá la página correspondiente a esa clase donde se listan todas las instancias reportadas en la literatura que están anotadas en ELM.

I- ¿En qué tipos de proteínas se encuentra el motivo LxCxE?

II- Existen dos tipos de instancias.

- True Positives (TP): Son instancias identificadas por la expresión regular y que la evidencia experimental muestra que es funcional.
- False Positives (FP): Son instancias identificadas por la expresión regular, que la evidencia experimental sugieren que son funcionales, pero cuando fue evaluada se cree que no es realmente funcional.

¿Se te ocurre algún ejemplo donde esto pueda ocurrir?

2. Busca en ELM E1A_ADECR.

a. ¿Cuál es el contexto taxonómico?
b. ¿Cuántas instancias anotadas hay? ¿Se encuentran los motivos anotados de E1A_ADE05? ¿A qué puede deberse?

3. La proteína Retinoblastoma (Rb) controla la transición en el ciclo celular de la fase G1 a la fase S mediante la interacción con factores de transcripción de la familia E2F.

a. Vaya a ProViz (http://proviz.ucd.ie/) y busque la proteína E2F1_HUMAN (Q01094).
b. ¿Puede identificar el motivo de interacción con Rb? (Pista: Hay una línea a la izquierda que se llama ELM).
c. El motivo ¿Está en un contexto estructural desordenado? ¿Se encuentra conservado?¿Es el mismo motivo usado por la proteína E1A para interactuar con Rb?
d. ¿Qué otros motivos identifica? ¿Algunos de estos motivos están involucrados en el ciclo celular?

Ejercicio Adicional 3. Familiarizándose con la base de datos ELM.

1. Pega y copia la siguiente secuencia en ELM y utiliza los parámetros que se indican a continuación.

```
> P12931
MGSNKSKPKDASQRRRSLEPAENVHGAGGGAFPASQTPSKPASADGHRGPSAAFAPAAAE
PKLFGGFNSSDTVTSPQRAGPLAGGVTTFVALYDYESRTETDLSFKKGERLQIVNNTEGD
WWLAHSLSTGQTGYIPSNYVAPSDSIQAEEWYFGKITRRESERLLLNAENPRGTFLVRES
ETTKGAYCLSVSDFDNAKGLNVKHYKIRKLDSGGFYITSRTQFNSLQQLVAYYSKHADGL
CHRLTTVCPTSKPQTQGLAKDAWEIPRESLRLEVKLGQGCFGEVWMGTWNGTTRVAIKTL
KPGTMSPEAFLQEAQVMKKLRHEKLVQLYAVVSEEPIYIVTEYMSKGSLLDFLKGETGKY
LRLPQLVDMAAQIASGMAYVERMNYVHRDLRAANILVGENLVCKVADFGLARLIEDNEYT
ARQGAKFPIKWTAPEAALYGRFTIKSDVWSFGILLTELTTKGRVPYPGMVNREVLDQVER
GYRMPCPPECPESLHDLMCQCWRKEPEERPTFEYLQAFLEDYFTSTEPQYQPGENL
```

Cell Compartment: **Not specified**
Motif Probability Cutoff: **100**
Context information: **(leave blank)**

a. ¿Cuántas instancias del motivo se encuentran?
b. ¿Que se puede decir sobre la estructura de la proteína? ¿Se observa algún dominio? ¿Se observan regiones desordenadas?

2. Realiza la misma búsqueda (P12931) utilizando los siguientes parámetros:

- Cell Compartment: **cytosol**
- Motif Probability Cutoff: **0.01**
- Context information: **Homo sapiens**

a. ¿Cuántas instancias aproximadamente se encuentran ahora?
b. ¿Cuántas instancias están anotadas?
c. ¿Los predictores estructurales y filtros (SMART, GlobPlot, IUPRED, Secondary Structure) coinciden sobre qué regiones son estructuradas/desordenadas?
d. Compara la ubicación de las instancias anotadas con la información estructural proveniente de IUPRED

3. Realiza la búsqueda de la secuencia de la proteína Paxillina (P49023) en ELM, utilizando los parámetros por defecto. Compara los resultados con una búsqueda de la misma secuencia pero modificando el parámetro **cellular compartment** 'plasma membrane'.

4. Busca la proteína SRC_MOUSE (P05480) en ELM.
   a. ¿Existen instancias anotadas?
   b. Si no, cuál es la instancia anotada más cercana que se puede encontrar. Investiga de dónde proviene esta información.

5. Busca la proteína CDN1A_HUMAN (P38936) en ELM.
   a. ¿Cuántas instancias de la clase DOC_PP1_RVXF_1 se encontraron?
   b. ¿Cuál es la diferencia entre estas instancias? ¿Que tiene de especial la instancia en la posición 155? ¿Porqué?

6. Busca la proteína 'P53_HUMAN'
   a. ¿Qué compartimentos celulares se le asignaron? ¿Tienen sentido?
   b. ¿Cuántos degrons hay en p53?
   c. ¿Existe algún sitio CDK en p53? ¿Hay un Cyclin Box en p53?

7. Busca en ELM la proteína 'MDM4_HUMAN' y encuentra el motivo de unión a USP (DOC_USP7_MATH_1). ¿Cuántas instancias del motivo se encuentran en esta secuencia?

8. Busca en ELM la proteína 'AMPH_HUMAN' y encuentra la clase 'LIG_Clathr_ClatBox_1'
   a. ¿Cuál es la relevancia biológica de cada una de estas instancias?
   b. ¿La anotación de la relevancia biológica coincide con la estructura globular?

9. Busca todas las instancias anotadas para "*Homo sapiens*" que contienen el término "cilium" (pista: Usa url http://elm.eu.org/elms/browse_instances.html).
   a. ¿Cuántas instancias hay?
   b. ¿Qué evidencia experimental está anotada y cuán confiable es esta evidencia?
   c. Descarga estas instancias en un archivo ".tsv" (tab-separated values)

10. Busca todas las instancias anotadas que contienen el término "retinoblastoma" (Pista: usa url http://elm.eu.org/elms/browse_instances.html)
   a. Compara el número de instancias humanas con el número de instancias virales.
   b. ¿Porqué hay tantas proteínas virales que interactúan con retinoblastoma? (Pista: La respuesta está en el abstract de la clase del motivo LIG_Rb_LxCxE_1)

Ejercicio Adicional 4. Proteína CagA de Helicobacter pylori.

La infección por *H. pylori* puede causar gastritis, úlcera péptica o cáncer de estómago. Hay una mayor probabilidad de desarrollar cáncer estomacal si la infección es producida por una cepa del Este asiático (como F32) en comparación a una cepa del Oeste (como NCTC 11637). Estas cepas difieren en el número y contexto de secuencia de los motivos EPIYA (Higashi, H., et al., 2002; Jones, K.R., et

al., 2009).

a. Copia y pega en ELM las secuencias la proteína CagA de una cepa del Oeste y una cepa del Este asiático, especificando Cytosol como compartimento celular, *Homo sapiens* como especie y un umbral de corte de la probabilidad del motivo de 0.001.

> NCTC11637_CagA

MTNETIDQQPQTEAAFNPQQFINNLQVAFLKVDNAVASYDPDQKPIVDKNDRDNRQAFDGISQLREEYSNKAI
KNPTKKNQYFSDFINKSNDLINKDNLIDIGSSIKSFQKFGTQRYRIFTSWVSHQNDPSKINTRSIRNFMENII
QPPIPDDKEKAEFLKSAKQSFAGIIIGNQIRTDQKFMGVFDEFLKERQEAEKNGEPTGGDWLDIFLSFVFNKE
QSSDVKEAINQEPVPHVQPDIATTTTHIQGLPPESRDLLDERGNFSKFTLGDMEMLDVEGVADIDPNYKFNQL
LIHNNALSSVLMGSHNGIEPEKVSLLYAGNGGFGAKHDWNATVGYKNQQGDNVATLINVHMKNGSGLVIAGGE
KGINNPSFCLYKEDQLTGSQRALSQEEIRNKIDFMEFLAQNNAKLDNLSEKEKEKFQNEIEDFQKDSKAYLDA
LGNDRIAFVSKKDPKHSALITEFGKGDLSYTLKDYGKKADRALDREKNVTLQGNLKHDSVMFVNYSNFKYTNA
SKSPDKGVGVTNGVSHLDAGFSKVAVFNLPDLNNLAITSFVRRNLENKLVTEGLSLQEANKLIKDFLSSNKEL
VGKALNFNKAVADAKNTGNYDEVKKAQKDLEKSLRKREHLEKEVEKKLESKSGNKNKMEAKAQANSQKDKIFA
LINKEANRDARAIAYSQNLKGIKRELSDKLEKINKDLKDFSKSFDEFKNGKNKDFSKAEETLKALKGSVKDLG
INPEWISKVENLNAALNEFKNGKNKDFSKVTQAKSDLENSVKDVIVNQKITDKVDNLNQAVSMAKATGDFSRV
EQALADLKNFSKEQLAQQTQKNESFNVGKKSEIYQSVKNGVNGTLVGNGLSGIEATALAKNFSDIKKELNEKF
KNFNNNNNNGLENEPIYAKVNKKKTGQVASPEEPIYAQVAKKVNAKIDRLNQAASGLGGVGQAGFPLKRHDKV
DDLSKVGRSVSPEPIYATIDDLGGPFPLKRHDKVDDLSKVGRSVSPEPIYATIDDLGGPFPLKRHDKVDDLSK
VGRSVSPEPIYATIDDLGGPFPLKRHDKVDDLSKVGLSRNQELAQKIDNLSQAVSEAKAGFFSNLEQTIDKLK
DSTKYNSVNLWVESAKKVPASLSAKLDNYATNSHTRINSNIQNGAINEKATGMLTQKNPEWLKLVNDKIVAHN
VGSVPLSEYDKIGFNQKNMKDYSDSFKFSTKLNNAVKDVKSSFTQFLANAFSTGYYSLARENAEHGIKNVNTK
GGFQKS

> F32_CagA

MTNETIDQTTTPDQTGFVPQRFINNLQVAFIKVDNAVASFDPDQKPIVDKNDKDNRQAYEKISQLREEYANKA
IKNPAKKNQYFSDFINKSNDLINKDNLIAVDSSVESFRKFGDQRYQIFTSWVSLQKDPSKINTQQIRNFMENV
IKPPISDDKEKAEFLRSAKQSFAGIIIGNQIRSDEKFMGVFDESLKARQEAEKNAEPAGGDWLDIFLSFVFNK
KQSSDLKETLNQEPRPDFEQNLATTTTDIQGLPPEARDLLDERGNFFKFTLGDVEMLDVEGVADKDPNYKFNQ
LLIHNNALSSMLMGSHSNIEPEKVSLLYGDNGGPEARHDWNATVGYKNQQGNNVATLINAHLNNGSGLIIAGN
EDGIKNPSFYLYKEDQLTGLKQALSQEEIQNKVDFMEFLAQNNAKLDNLSEKEKEKFQTEIENFQKDRKAYLD
ALGNDHIAFVSKKDPKHLALVTEFGNGELSYTLKDYGKKQDKALDGETKTTLQGSLKYDGVMFVNYSNFKYTN
ASKSPNKGLGTTNGVSHLEANFSKVAVFNLPNLNNLAITNYIRRDLEDKLWAKGLSPQEANKLIKDFLNSNKE
MVGKVSNFNKAVAEAKNTGNYDEVKKAQKDLEKSLRKREHLEKEVAKKLESRNDNKNRMEAKAQANSQKDKIF
ALISQEASKEARVATFDPYLKGVRSELSDKLENINKNLKDFGKSFDELKSGKNNDFSKAEETLKALKDSVKDL
GINPEWISKIENLNAALNDFKNGKNKDFSKVTQAKSDLENSIKDVIINQKITDKVDNLNQAVSEIKLTGDFSK
VEQALAELKNLSLDLGKNSDLQKSVKNGVNGTLVSNGLSKTEATTLTKNFSDIRKELNEKLFGNSNNNNNGLK
NNTEPIYAQVNKKKTGQATSPEEPIYAQVAKKVSAKIDQLNEATSAINRKIDRINKIASAGKGVGGFSGAGRS
ASPEPIYATIDFDEANQAGFPLRRSAAVNDLSKVGLSREQELTRRIGDLSQAVSEAKTGHFGNLEQKIDELKD
STKKNALKLWVESAKQVPTSLQAKLDNYATNSHTRINSNVQSGTINEKATGMLTQKNPEWLKLVNDKIVAHNV
GSAPLSAYDKIGFNQKNMKDYSDSFKFSTKLNNAVKDIKSSFVQFLTNTFSTGSYSLMKANVEHGVKNTNTKG
GFQKS

b. ¿Cuáles son las diferencias en las predicciones del motivo EPIYA? ¿Existen diferencias en la asignación por homología?

JalView

Ejercicio Adicional 1. Usando JalView con la proteína TIR aislada de *E. coli* patogénica

Las proteínas TIR son secretadas por la cepa patogénica de *E. coli* y se asocian a ciertas células de mamíferos, proyectando sus extremos N- y C-terminal a través de la membrana plasmática hacia la parte interior de la célula huésped tomando el control de la regulación celular local, por ejemplo induciendo junto con otras proteínas la formación de un pedestal de actina esencial para el ciclo patogénico de esta bacteria. La porción central de la proteína TIR permanece en el compartimiento extracelular y se asocia a la bacteria. Existen numerosas secuencias de TIR obtenidas de diferentes aislamientos de *E. Coli* patogénica almacenadas en la base de datos UNIPROT.

1. Carga el alineamiento de proteínas TIR que se encuentra en la carpeta MSA del TP de la materia (**tir_aligned.fasta**) en la ventana de JalView.
2. La expresión regular del motivo de unión a Ciclina es:

   ```
   [RK].L.{0,1}[FLMP]
   ```

   La expresión regular del motivo de fosforilación por CDK (quinasa dependiente de ciclina) es:

   ```
   [ST]P.{0,2}[RK]
   ```

   La fosforilación de proteínas durante el ciclo celular es realizada por complejos Ciclina-CDK, y requiere la presencia de ambos motivos en la proteína a ser fosforilada.

   Utiliza las expresiones regulares para encontrar estos motivos en las secuencias. Para poder resaltarlas, en la ventana donde ingresaste la expresión regular cliquea en *New Feature*. Ahí puedes crear un grupo y seleccionar un color para el mismo.
   a. ¿Todas las secuencias tienen ambos motivos?
   b. ¿Los distintos ejemplos de motivos están alineados o se encuentran en lugares diferentes de la secuencia?
   c. Algunos motivos están yuxtapuestos ¿consideras que pueden ser los dos funcionales al mismo tiempo?
3. Existe evidencia que el ciclo celular puede ser interrumpido por la cepa patogénica de *E. coli* (PMID: 11598051).

   El dominio SH2 une un motivo que posee una tirosina fosforilada. Busca el motivo SH2 utilizando la expresión regular del motivo SH2:

   ```
   Y..[IVLM]
   ```

   a. ¿Todas las secuencias tienen motivos SH2?
   b. ¿En base a tu respuesta anterior, esperas que las proteínas TIR sean o no

fosforiladas por tirosin quinasas dentro de la célula?

Ejercicio Adicional 2. Usando JalView con la proteína CagA aislada de la cepa *Helicobacter* patogénica

Las proteínas efectoras CagA son secretadas por la cepa patogénica de Helicobacter ingresando directamente al citoplasma de la célula huésped, en parte utilizando un motivo funcional llamado “motivo EPIYA”. Estas proteínas modulan el citoesqueleto de actina y el estado general de la célula.

1. Carga el alineamiento de las proteínas CagA que se encuentra en la carpeta MSA del TP de la materia (CagA_aligned.fasta). Busca el motivo EPIYA utilizando la expresión regular del motivo: `EP[IL]Y[TAG]`
    a. ¿Las secuencias de CagA: tienen un motivo EPIYA o más de uno?
    b. ¿Todas las secuencias tienen el mismo número?
    c. ¿Cuál es el mayor número de motivos EPIYA en una proteína?
    d. ¿Algún motivo EPIYA se superpone con algún motivo SH2?
    e. ¿Crees que las proteínas CagA son fosforiladas por tirosin-quinasas?