Análisis de alineamientos múltiples de secuencia de proteínas - Visualizando alineamientos con Jalview

Lucía B. Chemes, Juliana Glavina

Introducción

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

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

Guía de Ejercicios - JalView

El objetivo principal de estos ejercicios iniciales es:

- 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)

A. ¿Qué longitud tiene la proteína p53 humana?
B. Observar el esquema modular de p53: ¿Puedes identificar qué dominios PFAM tiene p53? ¿Qué nombres y qué funciones tienen?
C. ¿En qué regiones de la secuencia se encuentran estos dominios? Anotar de qué residuo a qué residuo abarca cada dominio, para usar más adelante.
D. ¿Creen que estos dominios corresponden unívocamente a dominios globulares? ¿A qué cree que corresponden las regiones marcadas como “Disorder” y “Low Complexity” en p53?

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 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 abre el archivo **p53_aligned.fasta**)

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 ¿porqué crees que es esto?
   b. ¿Todas las secuencias comienzan con el aminoácido metionina? A qué corresponden las secuencias que no?
   c. ¿Si quieren construír un alineamiento de alta calidad, preservarían o descartarían estas secuencias?
4. 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"
5. Es posible reorganizar el orden de las secuencias en el alineamiento, ordenando por ejemplo por identidad de secuencia. Para esto:
   Calculate → Sort → By Pairwise Identity.
   ¿Qué diferencia se observa con el alineamiento anterior?

ANÁLISIS DE 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.

6. 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?

7. 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%

   a. ¿Qué regiones espera que muestren una alta conservación en un MSA? Mirando el gráfico de conservación en la parte inferior de la ventana de Jalview, responda: ¿Qué regiones se encuentran más conservadas y cuales menos en p53?

Utilizando el filtro, respondan:

   b. ¿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?

   c. En base al análisis de conservación y al de identidad realizados: ¿qué regiones espera que correspondan a dominios globulares en p53?

      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:

d. ¿Qué tipos de patrones observa?

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

7. Identificación de motivos cortos de interacción:

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 la siguiente expresión regular:

`F...W.{2,3}[VIL]`

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?

8. Edición de secuencias

Vuelva a examinar el alineamiento.

a. ¿Existen regiones del alineamiento que no estén alineadas correctamente?
b. 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 5. 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 juxtapuestos ¿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 6. 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?